Calibration and Accuracy Testing of Weigh-in-Motion Systems

PETER DAVIES AND FRASER SOMMERVILLE

Examined in this paper are the problems of calibration and accuracy testing for weigh-in-motion systems, and several approaches are proposed that might improve the present situation. After the definition of terminology, recent calibration techniques are outlined and a different approach is advocated, based on random vehicles. A statistical appraisal of the approach concludes the first part of the paper. In subsequent sections, a new technique is described for self-calibration of weigh-in-motion systems, based on the fact that loads on certain axles of particular truck classes show relatively little variation. Self-calibration offers the prospect of improved weigh-in-motion accuracy in between conventional calibration exercises. In the final sections of the paper, the questions of weigh-in-motion accuracy appraisal and weigh-in-motion performance standards are addressed. An accuracy "funnel" for the assessment of weigh-in-motion performance is proposed. This provides for an absolute accuracy tolerance at low axle weights, with a percentage tolerance at higher loads.

Weigh-in-motion (WIM) systems have been used for many years in the collection of axle load data and as a screening tool for the identification of potentially overweight vehicles. As the use of WIM systems has increased, a number of issues have emerged concerning the operation and evaluation of these systems. The authors of this paper set out to examine the fundamental problems associated with calibration and accuracy testing of WIM systems, and propose several approaches to improve the current situation.

Much of the content of the paper has been developed through a contract being undertaken by Castle Rock Consultants on behalf of the states of Iowa and Minnesota. This contract is for the development and evaluation of a low-cost automatic weight and classification system (AWACS) based on a piezoelectric axle load sensor. As part of this work, a test program was developed to statistically assess the accuracy of the system over a period of time.

Before aspects of calibration and accuracy testing are described, the commonly used terminology is outlined. A review of current calibration techniques has been included, together with a discussion of the potential for self-calibration of WIM systems. Following this, the main emphasis of the paper is on statistical techniques to assess the accuracy of WIM equipment. At the conclusion of the paper the problem of WIM performance specifications is examined, advocating a new approach to accuracy assessment.

TERMINOLOGY

The technique of in-motion weighing attempts to measure the mass of a vehicle, a wheel, an axle, or a group of axles on a vehicle by determining the constantly changing vertical dynamic force that is applied by the tires to the pavement surface. The vehicle mass remains constant, by definition, but the rolling wheel applies a dynamic force to the highway that might range from double the static weight, after the vehicle has traversed a bump, to zero when the tire bounces off the ground. This so-called dynamic weight of an axle or vehicle makes up a component caused by the static load of the vehicle and a superimposed dynamic force component, which results from the fluctuating motions imparted to the vehicle by external factors. The magnitude of the dynamic force component is dependent on vehicle, highway, and environmental characteristics *(1)*.

The accurate measurement of static loads for vehicles at rest is surprisingly difficult. In order to weigh the whole of a truck in one instant, scales have to be large and complex. Placement of the vehicle on the scale can affect scale response significantly. Worse, however, is the fact that most scales weigh trucks section by section; each time the truck moves forward its suspension system shifts and redistributes load between the axles. Leveling the pavement and releasing the truck's brakes can reduce these static weight measurement errors; significant errors, however, inevitably remain.

To infer accurate static axle or vehicle loads from dynamic measurements, the vertical acceleration of all the vehicle elements should ideally be zero. The sum of the vertical forces exerted on a smooth, level surface by the perfectly round and dynamically balanced rolling wheels of a vehicle at constant speed in a vacuum is theoretically equal to the static weight of the vehicle. This ideal situation is obviously impossible to attain in practice. In any event, external variables, such as vehicle, highway, and environmental characteristics, are only some of the factors that influence the WIM measurement. Internal errors associated with the measuring equipment can also contribute significantly to discrepancies between dynamic weights and their static equivalents.

In most cases, weigh in motion can only give an instantaneous indication of the axle or wheel load as the vehicle crosses the WIM scale. If the WIM system were to operate without measurement error, the measured load would be the true dynamic force exerted by the wheel at a particular point and time. Users of weigh-in-motion systems, however, frequently want to infer static weights from dynamic measurements. The actual difference between static and instantaneous dynamic weight is often treated as an integral part of the WIM error. In common usage, therefore, WIM errors are made up of three components, whose order of importance might be *(a)*

Castle Rock Consultants, 5125 N. 40th St., #F119, Phoenix, Ariz. 85018.

actual static/dynamic force differences, *(b)* dynamic force measurement errors, and *(c)* static load measurement errors. It is generally impracticable to separate out these effects, and the remaining discussion treats these as a single, combined error or inaccuracy.

Terminology used to express so-called WIM accuracies varies. An impact factor (IF) can be calculated for axles or vehicles to give an indication of the system performance *(2)*

where

$$IF = \frac{\text{WIM weight}}{\text{Static weight}}$$

Alternatively, WIM accuracies can be expressed in terms of static-to-dynamic weight differences using either absolute or percentage weight values. Absolute differences would be appropriate if weighing errors were approximately equal, irrespective of vehicle or axle weight. Percent differences are more appropriate if the size of the weighing error increases in proportion to the mass of the axle being weighed, which is more likely the case. If neither of these conditions is true, separate accuracies should perhaps be quoted for axles in different weight bands. The percentage difference (PD) and absolute difference (AD) can be defined by

$$PD = \frac{\text{WIM weight} - \text{static weight}}{\text{Static weight}} \times 100\%$$

$$AD = \text{WIM weight} - \text{static weight}$$

Whether IF, PD, or AD is used to express WIM accuracy, results for a single vehicle are of little interest in themselves. Random fluctuations may make static/dynamic differences (errors) substantial, or negligible, for any particular vehicle or axle. Accuracy assessment requires WIM system performance to be established over large samples of vehicles, for which a normal distribution of errors will commonly be observed. Performance across the sample is described by statistical measures associated with two further error parameters: systematic error and random error. Both kinds of error can be associated with internal measurement errors, or actual static/dynamic weight discrepancies resulting from external highway factors.

The systematic error is given by the mean of the error distribution for individual measurements, whereas the random error is measured by its standard deviation. Systematic errors can arise for reasons relating to the design, installation, or operation of the system and cause a repeatable bias in all measurements carried out at a particular time. Random errors, however, are uncontrollable and unpredictable, and are intrinsic to any measurement. The purpose of calibration is to compensate for systematic errors, reducing them as far as possible. The initial calibration may change over time, producing a varying systematic bias in the load measurements.

Current WIM system calibration techniques are briefly reviewed in the next section. The feasibility of a self-calibration technique is described in the subsequent section.

CALIBRATION TECHNIQUE

Current techniques for calibrating weigh-in-motion systems vary considerably among practitioners. Little research has been performed to develop a standard technique, with present practice appearing to have evolved from trial and error. A brief review of some of the different techniques adopted is presented below.

Hamrick describes the technique adopted by Idaho Department of Transportation to calibrate a bending plate WIM system *(3)*. Essentially, it involved using a three-axle test vehicle of 30,000 lb gross weight making a series of runs over the systems of 20, 40, and 60 mph. The system calibration was then adjusted to minimize the average differences between the dynamic and static gross weights. Repeat calibration exercises were undertaken approximately every 3 months but no data are available indicating the extent of the change in systematic error over time.

The approach used by the Minnesota Department of Transportation involved three types of test vehicle *(4)*. A two-axle, six-tire vehicle was used to estimate the variations between dynamic and static weights and allow a preliminary calibration to be made. Subsequently, three vehicles of different weights were run over the weight sensor at different speeds, with sample sizes ranging from 4 to more than 70. The systematic errors for each test vehicle ranged between −2.2 percent and +4.7 percent. No further adjustments were made to the calibration. However, in subsequent calibration exercises several months later it was reported that the systematic errors now lay in the range −16 percent to +30 percent. No firm conclusions were drawn from these results, but it was suspected that the equipment was faulty.

Chow performed an evaluation of the PAT and Streeter Amet WIM systems in 1982 *(5)*. As part of the evaluation, a two-stage calibration procedure was adopted: a preliminary calibration wherein a loaded truck was driven across the scales at different speeds, at least 15 times at each speed, and the calibration settings adjusted to minimize the systematic error; followed by a subsequent main calibration using random trucks with various axle combinations, suspensions, and loadings, crossing at various speeds. The whole calibration procedure, according to the report, was "laborious, tedious and difficult, and required about six days to accomplish."

An alternative calibration method developed by the Transport and Road Research Laboratory (TRRL) in the United Kingdom is described by Priest and Moore *(6)*. This particular technique involves the application of a static load, via a load cell, using a hydraulic jack acting against the underside of a specially reinforced heavy vehicle. Loads are applied to the center of the weigh scale through a rubber-faced mild steel disc, 8 in. in diameter. However, such static calibration can take no account of systematic errors resulting from dynamic effects such as pavement profile. The use of a mobile dynamic loading rig for WIM calibration has also been suggested, but this would suffer from the same limitation.

To summarize, the use of test vehicles for calibration is not wholly satisfactory because the effect of factors such as suspension type is not taken into account. Mobile loading rigs take no account of pavement profile. The best approach to calibration, while costly, would seem to involve the use of random vehicles selected from the population of vehicles to be weighed at a particular site.

INITIAL CALIBRATION

A perfect calibration would eliminate all systematic error in weight measurement for the population of trucks at that site. It is not practicable to measure the mean of the population error distribution directly, because it is impossible to measure every vehicle at a site. However, this value may be estimated statistically using a random sample from the vehicle population of interest.

The systematic error for the calibration sample can be eliminated by setting the equipment so that the sample has zero mean error. In practice, this may be derived by plotting the WIM output against static load for the sample and fitting the best straight line through the points. The sample size for initial calibration depends on the calibration accuracy required and the inherent variability of the data.

Previous experience indicates that the standard deviation (SD) of the PD distribution will be around 10 percent. The standard error of the mean (SE_m) is given by

$$SE_m = \frac{SD}{(n)^{1/2}}$$

where n is the number of static/dynamic weight comparisons. Confidence limits of 95 percent are given by approximately ±2 SE_m. Therefore, for a calibration accurate to ±1 percent, with 95 percent confidence, we require

$$n = \left(\frac{10}{0.5}\right)^2 = 400 \text{ observations}$$

As each observation comprises one single, tandem, or triple axle static/dynamic comparison, approximately 150 trucks will be required to achieve the accuracy stated. Clearly, the sample size will vary for different confidence and accuracy levels.

SELF-CALIBRATION

Self-calibration, or automatic calibration, is an approach that might allow the systematic error to be minimized by continuously monitoring any change in the calibration. Self-calibration features might play a major role in reducing the operational costs currently associated with the above method or other conventional methods of WIM calibration.

The principle of the approach is that the loads on certain axles of specific truck classes show relatively little variation, regardless of the loading condition of the truck. Therefore, provided that the WIM system includes automatic vehicle classification, a data base can be built up by the system consisting of axle load measurements for these particular axles. If the mean weight of these axles is calculated after the addition of each new set of measurements to the data base, the system calibration factor can be adjusted to force the mean weight to agree with a known long-term population mean.

One particular axle category has been suggested as suitable for potential use as the basis for a calibration feature *(7)* (C. Dahlin, in a letter addressed to Perry Kent, FHWA, 1983) and *(8)*. This is the steering axle of 3S2 trucks. The vehicle dimensions of 3S2s are such that the kingpin is usually located close to the center of the first tandem axle. Therefore, the loading on the vehicle has relatively little effect on the steering axle load.

To test this hypothesis, a data base of U.S. vehicle dimensions and weights, collected during a biennial Truck Weight Study (TWS) in Arizona, was analyzed by the authors. For the self-calibration technique to be practical, only classes of vehicle that commonly occur in the normal traffic stream can be used, and so for this analysis only 3S2s were selected. Of the 1,500 vehicle entries in the data base, over one-third were of the class being analyzed. The vehicle weights in the data base had been obtained from portable static weighing scales, accurate to ±2 percent, and collected during a period when no enforcement weighing was in progress.

A computer program was written to analyze the data base by first selecting only the vehicle type under consideration and then examining the axle weight of the steering axle. Each of the weights examined was sorted into classes of 400 lb in the range 6,400 lb to 14,000 lb (Figure 1). In total, 512 vehicles were used in the analysis, having a mean axle weight of 9,950 lb, with an associated standard deviation of 1,126 lb.

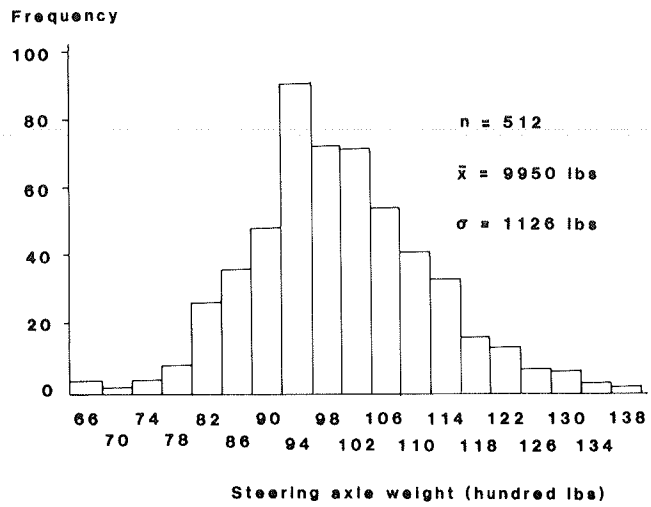

FIGURE 1 Analyzed 3S2 truck data.

If the assumption is made that the standard deviation of the sample is representative of the population as a whole, it is possible to determine the sample size required to give a specified standard error of the mean for a given confidence level. For example, for 95 percent confidence limits of ±100 lb in 10,000 or approximately ±1 percent, the required sample is given by

$$SE_m \text{ (standard error of mean)} = \frac{\sigma}{(n)^{1/2}}$$

where

σ = standard deviation and
n = sample size.

95 percent confidence = 1.96 SE_m (normally taken as 2 SE_m)

Therefore

$$1.96 \times 50 = \frac{1{,}126}{(n)^{1/2}}$$

where

$n = 132$

From the analysis of the limited data available, it would appear that the use of steering axle weight data from 3S2s for self-calibration purposes would be a practical proposition, although a number of issues would first need to be resolved. In particular, the reliability of the data used to determine the population mean and standard deviation would need to be assessed. It is possible that the data employed for this analysis are biased, caused by a proportion of illegally overloaded vehicles bypassing the weighing location even though no citations were being brought. These data may also be site dependent and time dependent as the construction and use of the vehicles change. This could be readily investigated given appropriate data for different sites and years, but would need to be monitored in the future. The accuracy of the vehicle classification, particularly for the class used in this technique, is a further consideration for the method to be effective and would need careful examination.

Because of the problems above, the technique would not totally eliminate the need for individual site calibration on a regular basis but could be used in several different ways to improve the accuracy of axle weight measurements between calibrations. For example, successive sample means could be calculated in turn, or a long-term moving average of the samples could be derived. When there was a significant difference between the sample means and the population mean, the system could just record the fact and do nothing, or it could alert operating personnel that the equipment was in need of recalibration. Alternatively, the system could automatically readjust the calibration factor by a small increment in the required direction. These fluctuations in the calibration factor could go unrecorded, or, more likely, would be recorded so that the weight data could subsequently be "unadjusted" should the need arise.

The concept of self-calibration appears to offer the potential for improved long-term accuracy of WIM systems. However, before the technique can be fully implemented, further detailed analysis of weight and classification data will be necessary, together with practical testing in the field to prove the concept. Castle Rock Consultants is currently addressing this need under contract to the states of Iowa and Minnesota.

ACCURACY EVALUATION

The accuracy of WIM systems can be directly determined by obtaining values for the mean and standard deviations of the PD or AD distributions, for both individual axle and gross vehicle weights, at various points in time following the initial system calibration. As with the initial calibration, it is impossible to measure the standard deviations directly, because not every vehicle that crosses the site can be checked. Consequently, a sample must again be taken, and the standard deviation of the sample used to obtain an unbiased estimate of the standard deviation of the population. It is possible to estimate confidence limits on this estimate of the population standard deviation. According to Spiegel *(9)*, the SE of this standard deviation is given by

$$\mathrm{SE} = \frac{\mathrm{SD}}{(2n)^{1/2}}$$

Therefore, for an initial calibration using 150 trucks (400 observations), 95 percent confidence limits on the standard deviation would be about ± 0.7 percent.

Each sample can be divided into weight ranges to detect differences in errors between weight ranges and statistical tests performed to identify significant differences between ranges.

Following the initial calibration, a second sample of weight measurements may be taken to observe whether any significant change in calibration or accuracy has occurred. A change in calibration will lead to different sample means, and changes in random scatter will lead to different sample variances. The two-tailed Students t-test can be used to determine, at any desired level of confidence, whether the means of two samples drawn from the same population are significantly different.

In the t-test, the null hypothesis is that the two means are equal and any actual numerical difference is due only to that amount of scatter expected when sampling a normal distribution. The t value is defined by

$$t = \frac{X_1 - X_2}{S_d}$$

where X_1 = sample mean of the calibration sample and X_2 = sample mean of the second sample. S_d is defined by

$$S_d = \left[\frac{(N_1 - 1)\, S_1^2 + (N_2 - 1)\, S_2^2}{N_1 + N_2 - 2}\right]^{1/2} \left[\frac{1}{N_1} + \frac{1}{N_2}\right]^{1/2}$$

where N_1, N_2 are the sample sizes, and S_1^2, S_2^2 are the sample variances.

The sample size N_2 can be estimated for particular significant differences in the means and for particular confidence levels. If a difference greater than ±2 percent is required at 95 percent confidence levels, for example, the sample size N_2 needed can be estimated as follows:

Assume N_1 (original calibration) = 400 axles and that the standard deviations of the samples S_1, S_2 are both 10 percent, then

$$S_d = 10\% \left(\frac{1}{400} + \frac{1}{N_2}\right)^{1/2}$$

If detecting differences in the means at a 0.05 level of significance is desired, then statistical tables show that the critical values of t are ± 1.96 in this two-tailed test. That is, if $t > 1.96$ or $t < -1.96$, the difference in the means was not due to the random scatter expected when sampling a normal distribution but was because the population of the two samples were different. In this case it would mean that the calibration had changed, leading to a systematic bias.

Using these values to find the minimum number for N_2 gives

$$t = \frac{2\%}{S_d} = 1.96, \text{ or } S_d = \frac{2\%}{1.96}$$

As it has already been estimated that $S_d = 10\%\ [(1/400) + (1/N_2)]^{1/2}$, then solving for N_2 yields $N_2 = 126$ axles. This

number of axles corresponds to about 50 trucks. Therefore, a second sample of 50 random trucks should be sufficient to detect a 2 percent change in the calibration, with 95 percent confidence.

PERFORMANCE STANDARDS

With the variety of WIM systems currently available, different accuracy levels can be obtained. Low-speed systems are capable of high-accuracy levels, whereas high-speed systems are relatively less accurate. As indicated earlier, users frequently want to infer static weights from dynamic weights so that the difference between the two measured weights is often treated as an integral part of the WIM system error. Despite the influence of highway and vehicle characteristics, it is generally recognized that there should be WIM performance specifications within which the WIM systems should operate.

This issue is currently being addressed by a research contract being undertaken as part of the Heavy-Vehicle Electronic License Plate (HELP) program, for which Castle Rock Consultants is a management consultant. The Development of Weigh-in-Motion Performance Specification contract involves the testing of different WIM systems in the laboratory and field, and the definition of performance criteria including accuracy, durability, and reliability. These criteria will be defined as recommended levels of performance and will be a quantitative assessment of the characteristics of WIM systems that could be used in the HELP program.

A possible approach for accuracy specification has been proposed by the authors. The approach would involve the specification of different accuracy limits according to the axle weight. At lower axle weights, below 10,000 lb for example, the accuracy is probably best specified as an absolute value, whereas beyond this threshold a percentage is more appropriate. This would lead a funnel within which the measurements should lie. An absolute tolerance of 1,000 lb and a percentage tolerance of 10 percent above 100,000 lb might be appropriate values (Figure 2). The percentage of measurements falling within the funnel would provide an indication of the system's ability to meet the accuracy performance specification.

Although this approach does not eliminate the influence of highway, vehicle, and environmental factors, it does appear to offer a suitable measure of operating performance. Further work would be necessary to establish acceptable values for the funnel dimensions.

CONCLUSIONS

Outlined in this paper are current WIM calibration techniques and the approach of self-calibration is developed as an aid to improved accuracies between regular calibration exercises. Analysis of TWS data has demonstrated the potential of the self-calibration concept, which relies on the fact that loads on certain axles of particular truck classes show relatively little variation, regardless of loading condition. Currently, work is in progress to implement self-calibration of WIM in Iowa and Minnesota.

The statistical design of calibration and accuracy evaluations has also been investigated. In particular, techniques for determining the sample sizes of measurements required to achieve specified accuracy and confidence levels have been outlined. Using such techniques, statistically valid evaluations can be readily performed to assess the calibration and accuracy of WIM systems.

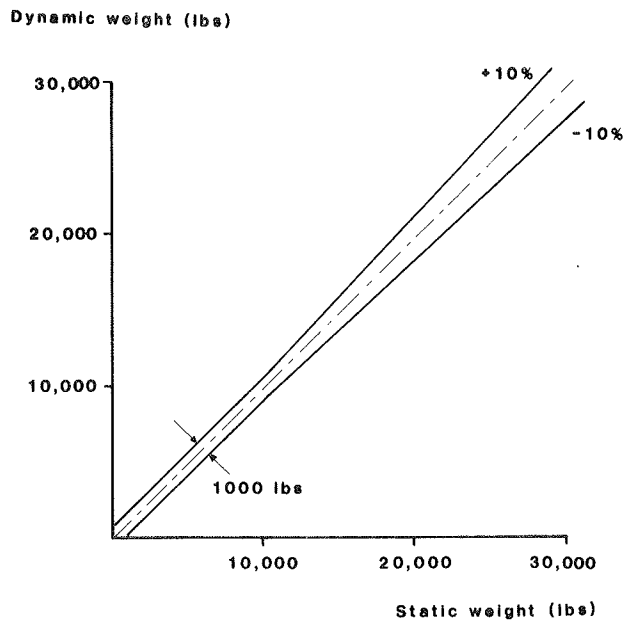

FIGURE 2 WIM performance specification.

An approach to the definition of a performance specification, or standard, has been proposed as part of this paper. At low axle weights, an absolute accuracy tolerance would seem to be appropriate, with a percentage tolerance for higher axle weights. This approach, producing a funnel-shaped tolerance band, may allow WIM systems to be specified and compared in similar ways.

REFERENCES

1. C. E. Lee. Concepts, Advantages and Applications of Weigh-in-Motion Systems. *Proc., 1st National Weigh-in-Motion Conference*, Denver, Colo., July 1983.
2. R. C. Moore. High-Speed Axle Load Measurement—A European Perspective. *Proc., 2nd National Weigh-in-Motion Conference*. Atlanta, Ga., May 1985.
3. J. Hamrick. State Experiences with Weigh-in-Motion Systems—PAT System. *Proc., 1st National Weigh-in-Motion Conference*. Denver, Colo., July 1983.
4. J. L. Wright. Status of Minnesota Department of Transportation's Weigh-in-Motion Program. Presented at the 62nd Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C., 1983.
5. W. Chow. *Evaluation of the PAT and Streeter Amet Weigh-in-Motion Systems*. California Department of Transportation, Sacramento, 1982.
6. R. A. F. Priest and R. C. Moore. *Equipment and Techniques for Commercial Vehicle Axle Load Surveys in the United Kingdom. The Estimation of Traffic for Road Pavement Design*. Transport and Road Research Laboratory. Crowthorne, England, 1981.
7. C. Dahlin. Minnesota's Experience with the IRD WIM System. *Proc, 1st National Weigh-in-Motion Conference*, Denver, Colo., 1983.
8. F. K. Sommerville. *Dynamic Axle Load Measurement*. Ph.D. dissertation, University of Nottingham, England, 1986.
9. M. R. Spiegel. *Theory and Problems of Statistics*. Schaum's Outline Series, New York, 1961.

Publication of this paper sponsored by Task Force on Weigh-in-Motion.