Comparative analysis of repeatability and reproducibility of compaction testing *(abridged for POMS proceedings)*

K.W McLain
Missouri Department of Transportation
D. P. Bumblauskas
University of Northern Iowa
Email: daniel.bumblauskas@uni.edu
D. J. White
Iowa State University
D. D. Gransberg
Iowa State University

Abstract
This article investigates possible alternatives for soil compaction testing to potentially replace the standard testing equipment by examining all the testing methods for reproducibility and repeatability. This study found that coefficients of variation and standard error to the mean produced the best results that fit with previous studies.

Keywords: Gauge R&R, Field operations, management decision making

INTRODUCTION AND RESEARCH MOTIVATION

The standard method for testing construction subgrade compaction for the Missouri Department of Transportation (MoDOT) is the American Association of State Transportation Officials (AASHTO, 2013) test method T310, *In Place Density and Moisture Content of Soil and Soil-Aggregate by Nuclear Methods (Shallow Depth).* The nuclear density gauge (NDG) has been the standard instrument for compaction testing since the late 1970's. The reliability of the nuclear density gauge (NDG) and the ability to determine gravimetric moisture content have made routine compaction testing quick and straightforward for field inspectors. However, the speed and convenience of the NDG comes with a price which includes inspector licensure with the federal government and required safety training, as well as special storage, transport, and field security procedures.

The Construction and Materials Division of MoDOT is actively investigating alternatives guided by the cost and time of the processes required when using the NDG and the

AASHTOWare™ Pavement ME Design, where mechanistic-empirical (ME) design procedures no longer are based on density-moisture content requirements. Pavement design is increasingly emphasizing the importance of achievement of minimum subgrade/base modulus rather the density-moisture alone.

The investigative process described in this article looks at selected alternative testing equipment and methods. Price, portability, testing time, ease of use, calibration requirements, accuracy, repeatability and reproducibility were parameters considered in comparing the various tests. To have confidence in a method and avoid conflicts between owner and contractor, the equipment and its associated testing protocol need to be accurate, repeatable and reproducible from operator to operator (gauge repeatability and reproducibility or GRR). While GRR has been used extensively in other applications, such as production manufacturing, quality control, and process improvement, the technique has rarely been used in field soil compaction applications. Therefore the objective of this research is to apply GRR in the comparative analysis of compaction testing devices and include its output in the decision process for choosing viable alternatives to the NDG.

Tested devices for this study included the Zorn ZFG 2000 Light Weight Deflectometer (LWD), and Dynamic Cone Penetrometer (DCP). The investigation of repeatability and reproducibility was done in the field on active construction sites rather than in the laboratory with technician-prepared soil filled drums/tubs or test strips (Mazari et al. 2013).

LITERATURE REVIEW

Literature searches for assessing repeatability and reproducibility for soil compaction testing went from specific searches to wider more broad searches on the subject of repeatability and reproducibility (R&R). The Automotive Industry Action Group (AIAG) Measurement System Analysis (MSA) Manual is the standard on defining R&R and for calculating R&R for parts and devices and was the starting point for the literature review. The AIAG MSA Manual contained preparatory background, guidelines, and computation processes for R&R range method, average and range method and analysis of variance (ANOVA) method. Mazari et al. (2013) reported on measuring repeatability and reproducibility of modulus measurement devices on prepared soil samples using the ANOVA and average and range methods. The AIAG manual and Mazari article led to branching out to other feasible statistical methods for measuring reproducibility and repeatability of compaction measuring devices.

Joubert and Meintjes (2015) employed AIAG average and range methods to examine the GRR of GPS data used by freight shippers., but also used the “Honest Gauge” R&R method proposed by Wheeler (2009). Wheeler recommended a system in which the percentage sum of the components of measurement total 100 percent. In depth comparisons of the AIAG described methods and the “Honest Gauge” methods were covered in a PhD dissertation (Stamm 2013) and in a Master’s thesis (Pandiripalli 2010).

Dhawale and Raut (2013) used one-way and two-way ANOVA calculated with a computer program to check GRR on tools, parts, operators and equipment. Seltman (2015) provided background on the workings and calculations for one-way ANOVA. ANOVA calculations led to the investigation of hypothesis testing statistics for determining GRR for varying compaction testing equipment. Mann (2010) and Gertsman (2006) have provided a thorough background on hypothesis testing and also provided good visualization of the method. Interpretation of p-values from the hypothesis testing was done by Fay and Gerow (2013) who gave context on p-values and the standard use of 0.05. Cowles and Davis (1982) and Nuzzo (2014), related the origins of using a level of significance of 0.05 and contemplated whether the use 0.05 value is truly correct.

Framework for information on coefficients of variation and standard error to the mean were derived from IDRE (2015), Fay and Gerow (2013), and Montgomery, Runger, and Hubele (2007). To have a point of comparison for coefficients of variations for modulus reporting compaction devices, White et al. (2009), Nazzal at al. (2007) and Alshibli, Abu-Farsakh, and Seyman (2005) were examined. MSHD (1975) was studied to possibly locate calculated standard deviations of the then new nuclear density gauge to the then standard sand cone and volume measure.

METHODOLOGY

Field Testing Procedure

Four sites were used to assess repeatability and reproducibility of measurements for the, LWD, and DCP. Tests of the three alternatives and the NDG were conducted on the following four construction project structural fills with the details on testing procedures and soil types and photographs of the testing devices are available in McLain (2015).

Prior to testing, the test locations were smoothed out using a hand shovel or a nuclear density gauge scraper plate. A nuclear density gauge (NDG) is used to produce two differing test areas and also a point of comparison. A nuclear gauge reading is taken and then the gauge is turned 180 degrees and a subsequent reading is taken. In the limits of the outline of the nuclear gauge test, five DCP readings per two testers are taken approximately three inches apart (McLain, 2015). This procedure usually limits the number of testers to two. In the second NDG test area five test trials of the LWD per tester are conducted, with the first tester performing the seating blow. This article reports the testing results from the LWD and DCP devices conducted at the Discovery Parkway project located just south of Columbia, Missouri.

AIAG Method

The AIAG Methods are defined by the Measurement System Analysis (MSA) Manual (4th edition). The MSA manual covers three different methods of analysis

- The Range method
- The Average and Range (A&R) method
- Analysis of Variance (ANOVA) method

The authors used the A&R method to investigate compaction test devices. The A&R method can estimate both repeatability and reproducibility with differing parts' role in the precision error of measurement. The A&R method can also estimate total precision error of measurement. This method allows for differing parts to be measured by several operators with several trials. The differing soil locations are the differing parts and are being measured by the compaction test devices several times with different operators. The A&R Method, however, does not consider the operator and device interaction. For MSA measurement and calculations please refer to the AIAG 2010 Manual and McLain (2015).

Wheeler's HG Method

Wheeler (2009) proposed an alternate to the AIAG GRR method, which he called "an honest GRR study". It is designated as the HG Method in this article. The HG Method differs from the AIAG method in that the sum of the components of measurement equals the Total Variation. For component Calculation please refer to Wheeler (2009) and McLain (2015).

The question that arises is whether the AIAG GRR and HG GRR are accurate measurement systems for determining repeatability and reproducibility of the DCP and LWD compaction testing methods. The MSA manual furnishes general GRR criteria guidelines as shown in McLain (2015).

Coefficient of Variation and Standard Error to the Mean

Coefficients of Variation (COV) of the results were calculated from the trials performed by the two differing operators. The COV is defined as the ratio of the standard deviation to the mean, with detailed equations provided by McLain (2015). The COV is useful because it is dimensionless and measurements using other units and differing means can be compared. In contrast, standard deviations themselves are in the context of the measured data and cannot be effectively compared to data with differing units. The standard error (SE) is the standard deviation of a sampling distribution (Montgomery et al. 2007).

Testing results were also analyzed using the statistical method of one-way ANOVA. The one-way ANOVA compares the means of data from differing groups (aka two differing operators performing compaction tests). The ANOVA statistic tests the null hypothesis. For one way ANOVA the general assumptions are normality, equal variance and independence of errors (Seltman 2015). For calculation and of null hypothesis and alternate hypothesis with supplementary calculations for F-Statistic please refer to Seltman (2015) and McLain (2015).

Generally, F-Statistics are near 1.0 when the null hypothesis is true and usually larger when the alternative hypothesis is true. The F-statistic can be compared to the F-critical. If the F-statistic is less than F-critical then the null hypothesis is thought to be true. Also the p-value can be compared to the alpha value or significance level, usually 0.05 (Cowles & Davis, 1982). To keep the null hypothesis, the p-value must be larger than α. The authors used commercially available programs to perform the one-way ANOVA statistical tests.

Hypothesis Testing

Like one-way ANOVA, hypothesis testing statistics requires that the null hypothesis has no significant difference between the means of groups or testers as seen in equation 19. The alternative hypothesis states that the test means are significantly different as presented in equation 20. The test statistic calculations are provided in McLain (2015).

The z_{stat} divides the area under a normal distribution curve into rejection and nonrejection regions for the null hypothesis. From the z_{stat}, a p-value is calculated. This is easily done using a statistical computer program. The p-value provides support against or for keeping the null hypothesis. The p-value is compared against a threshold value called the level of significance or alpha (α).The authors have applied the typical statistical convention (Gertsman 2006) as shown in McLain (2015).

FINDINGS AND RESULTS

The DCP readings from the Discovery Parkway project in blows per Inch (BPI) for 8+ inches of penetration for five separate trials for the 10 sites are presented below in Table 1for testers A and B.

Table 1. DCP Results Discovery Parkway

	Tester A Average of 5 Trials BPI	Tester B Average of 5 Trials BPI
Site 1	**0.3338**	**0.3336**
Site 2	**0.3767**	**0.2407**
Site 3	**0.3237**	**0.2872**
Site 4	**0.3104**	**0.2861**
Site 5	**0.3261**	**0.2907**
Site 6	**0.2897**	**0.2724**
Site 7	**0.2977**	**0.2770**
Site 8	**0.3392**	**0.2915**
Site 9	**0.2819**	**0.2734**
Site 10	**0.2932**	**0.2852**

The results are for five trials of 10 sites on the Discovery Parkway project site. The Zorn LWD results come in two forms: dynamic deflection modulus, (E_{vd}) in mega-newtons per squared meters, (MN/m^2), and settlement (s) in millimeters (mm) as shown in Table 2.

Table 2. LWD results Discovery Parkway

Test Site #	**Tester A Average of 5 Trials E_{vd} (MN/m^2)**	**Tester B Average of 5 Trials E_{vd} (MN/m^2)**
Site 1	**4.40**	**4.60**

Site 2	**3.92**	**4.10**
Site 3	**3.98**	**4.04**
Site 4	**3.98**	**4.30**
Site 5	**3.86**	**4.04**
Site 6	**3.62**	**3.96**
Site 7	**3.40**	**3.82**
Site 8	**3.82**	**4.02**
Site 9	**3.80**	**3.98**
Site 10	**3.26**	**3.70**

The AIAG and HG reproducibility and repeatability measurement results are provided in McLain (2015). For both the AIAG and HG methods the % GRR (Repeatability and Reproducibility) exceeds the 30 percent failure threshold of acceptability. These considerable results were not unanticipated given the testing was conducted on soil using standard construction compaction techniques employing heavy equipment.

The AIAG protocol states that if the range for an individual trial exceeds the calculated Upper Control Limit (UCL_R) for range for the entirety of the trials, that that trial(s) be redone or discarded and the upper control limit be recalculated for the remaining trials.

For the 10 sites a UCLR of 0.1507 was calculated. For tester A, Sites 2, 3, and 8 ranges met or exceeded the Upper Control Limit, and using AIAG protocol, Sites 2, 3 and 8 were removed for both Testers, giving seven remaining sites for which to evaluate AIAG and HG Gauge R&R. This is shown in Figure 1.

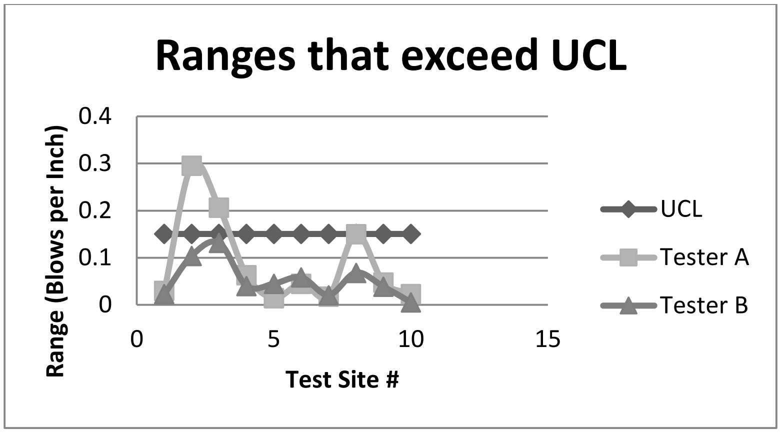

Figure 1. Ranges Exceeding Upper Control Limits for DCP Gauge R&R

Taking Sites 2, 3, and 8 out of the calculation lowered GRR by about 23 percent points but the figure was still over the maximum acceptance level by about 38 percentage points. The parts variation increased from 10 site set-up to the 7 site scenario because of the decrease in the number of parts (soil sites) where the K_3 constant increased. The LWD AIAG and HG gauge R&R 10 Site tests displayed similar results for the DCP 7 Site results with the LWD GRR being 10 percent lower than the AIAG accepted DCP results. The GRR results still exceeded AIAG standards for an acceptable system. The comprehensive results are provided in McLain (2015). Mazari et al. (2013), in laboratory conditions found for Zorn LWD, referred as a Portable Impulse Plate Load Device (PIPLD), the following Average and Range method results: EV% - Repeatability = 1; AV% - Reproducibility = 0.1; R&R = 1; % PV –Parts Variation = 99.The good repeatability and reproducibility results were from the research team rigidly controlling the moisture content and density of the soil being tested. The standard deviation of moisture content for all prepared lifts and specimens were 0.5 % with a range of 0.9% . The mean moisture content was 0.1% below optimum moisture content OMC.

The optimum moisture content for the Discovery Parkway site was 15.5%. Nuclear gauge moisture measurement for the 10 LWD subsites averaged 15.66 % with a range of 3 percent and standard deviation of 0.92%. The maximum dry density for the site was reported at 111.5 pcf. The average nuclear gauge readings for dry density of the 10 subsites were 106.7 pcf. The standard deviation for the site was 2.26 pcf with a range of 7.5 pcf.

COVs were calculated for the DCP and LWD as provided in McLain (2015). The COV values improved by about a factor of two when the three sites 2, 3, and 8 were removed through the AIAG conventions due to the large ranges encumbered by Tester A. White et al. (2009) conveyed COVs for DCP of 20% to 32%, measuring 12 in. deep on test strips. White et al. (2009), also reported COVs ranging from 29% to 61% for Zorn LWDs tested in cohesive to granular subgrades. Prima 100 LWDs (Alshibli et al. 2005) tested in laboratory conditions had COVs that ranged from 1.2 % in clay to 55.8% percent in sands, but for eight clay samples (soils like that found on Discovery Parkway site) the average COV was 18.2% . Nazzal et al. (2007) reported Prima 100 LWD COV results that varied from 2.1% to 28.1% for various highway construction bases and subgrades. It was noted that COV value decreased as the LWD elastic moduli increased.

Removing the outliers, (sites with test ranges outside AIAG specifications) decreased the COV for tester A approximately 3.5 % and the combined COV by about 3 percent. The COVs calculated for the DCP were lower than found in White et al. (2009), but more variation can be expected in DCP tests conducted in granular subgrades. The COVs for the tested Zorn LWD trended on the lower end when compared to the Prima 100 LWDs, but were in the range of reported results.

In attempting to take the soil variation from the entire site out and give an indication of reproducibility between testers , COV's for individual LWD trials were calculated and compared, as shown in Table 3. Percent change from the average varied as little of 1.23 percent to 119 percent.

Table 3. LWD COV for Individual Test Sites

Trial No.	Tester A COV	TESTER B COV	Average COV	Difference in COV	Percent Change from Avg. COV
Trial 1	0.0407	0.0238	0.0323	0.0169	52.40

Trial 2	0.0338	0.0154	0.0246	0.0184	74.80
Trial 3	0.0101	0.0198	0.0150	0.0097	64.88
Trial 4	0.0582	0.0147	0.0365	0.0435	119.34
Trial 5	0.04209	0.037	0.0395	0.0051	12.87
Trial 6	0.0506	0.0342	0.0424	0.0164	38.68
Trial 7	0.0372	0.0256	0.0314	0.0116	36.94
Trial 8	0.0305	0.0186	0.0246	0.0119	48.47
Trial 9	0.0166	0.0246	0.0206	0.0080	38.83
Trial 10	0.0245	0.0242	0.0244	0.0003	1.23

The Standard Error in percent of averages of around 2 percent was calculated for the DCP and LWD as provided in McLain (2015). They show a good accurate point of estimate for both average blows per inch for 8 inch depth for the DCP and modulus readings with the Zorn LWD for the 10 sites on the Discovery Parkway Site.

One way ANOVA and Hypothesis test results for paired samples are displayed below. The statistical methods can be used to look at the reproducibility of each tester. The p- values for both methods were generated by commercially available software. The methods differ as to what significance level to reject or fail to reject the null hypothesis that the difference between the means of the test results conducted by the two different testers are essentially equal. McLain (2015) summaries the hypothesis test results for this study.

CONCLUSIONS AND FUTURE WORK

For the MoDOT personnel and partnering contractors looking at the systems, the most understandable and useful statistics are the Coefficient of Variation and the Standard Error in Percent of Average. The COV is also a useful comparative element since it is unit-less; this allows for comparison among the differing testing devices that produce dissimilar test results. The key in understanding the concept of COV is the test data with the smaller COV is less dispersed than the variable with the larger COV (IDRI 2015). In a field test comparing the ten differing sites, the COV displays the amount of variation in the sites. Individual site COV’s show the variability between testers. The LWD Tests for each of the 10 sites were conducted with Tester A performing the initial three seating blows then conducting 5 sets of three drops. They recorded the average dynamic modulus and settlement after each three drops. Then Tester B repeated the process excluding the initial seating blows. McLain (2015) shows that Tester B had higher average modulus readings than Tester A. This would indicate that after the initial three seating blows that the soils of the Discovery Parkway project were still being compacted from Tester A drops. The difference in average modulus measurement ranged from 0.06 MPa to 0.44 MPa with the average difference being about 0.254MPa. A stiffer soil site would have displayed less variation and given a better indication of repeatability and reproducibility especially within individual trials.

The DCP Tests were not truly repeatable tests because the test is a destructive test and the distinct soil columns were obliterated. The test had to be averaged over the outline of the initial nuclear density test. This procedure introduced further variability into the measurements. The soil and degree of compaction on the overall test sites varied under concentrated testing terms but was fairly uniform in standard construction procedures.

The AIAG GRR method's stated thresholds or limits are subjective and there is no support for the limits in the MSA manual (Wheeler 2009). When Equipment Variability (repeatability) is found to be greater than Appraiser Variation (Reproducibility) as seen in the DCP tests, the probable causes are that the gauge needs to be repaired or replaced or there is excessive 'within part variation' (Pandiripalli 2010). The excessive within part variation is likely for the DCP tests in which every trial for both testers was an individual test in a varying medium (soil on a project). For Reproducibility greater than Repeatability, appraisers or operators need better training or the testing equipment needs to be recalibrated. In the case of LWD testing the part (soil) was changed in the testing process by becoming more dense, producing a higher modulus.

The AIAG and HG methods are designed more for manufactured parts or laboratory prepared specimens. The gauge R&R tests for the LWD would provide more consistent results if conducted on manufactured plates or on varying stiffness rubber pads (White et al. 2009). The AIAG and HG method require the removal or replacement of data if range of measurement between trials exceeds a calculated Upper Control Limit. If the tests were conducted correctly, that data has value and has significance. It can mean variation in the soil or a malfunction in the instrument and should be investigated as real data or an anomaly before removal from a data set.

Other challenges include the One Way ANOVA and Hypothesis Test for Paired Samples which are often not thoroughly understood by construction personnel without previous research, work experience or subject matter expertise. Secondly, the two tests are not definitive tests (Nuzzo 2014). When Ron Fisher introduced the concept of the P value in the 1920's, he envisioned it to be an informal method to determine if the data produced results that warranted further examination. Fisher intended the P value to part of a process that used both data and background knowledge to point to a scientific conclusion (Nuzzo 2014). The level of confidence is an additional query for field testing. The P value condenses data from a null hypothesis; it cannot indicate the basis for the data. The decision maker needs to have sufficient background on the data. The alpha value at 0.05 has become the standard and has been accepted by researchers as statistically significant or noteworthy. There are no guidelines as to what alpha value/ level of confidence to use when investigating field data versus lab data. This is a decision for the tester or other informed decision maker.

Bibliography

Alshibli, K.A., Abu-Farsakh, M. and Seyman, E. (2005). "Laboratory evaluation of the geogauge and light falling weight deflectometer as construction tools." *Journal of Materials in Civil Engineering*, 17(5), 560-569.

AASHTO (2014). "Density of Soil In-Place by the Sand Cone Method." T191, AASHTO, Washington D.C.

AASHTO (2013). "In Place Density and Moisture Content of Soil and Soil-Aggregate by Nuclear Methods (Shallow Depth)." T310, AASHTO, Washington D.C.

AASHTOWare version 2.2 (2015), (computer software), AASHTO, Washington D.C.

ASTM (2009). “Standard Test Method for Density of Use of the Dynamic Cone Penetrometer in Shallow Pavement Applications.” D6951/D6951M-10, ASTM International, West Conshohocken, PA.

ASTM (2011) “Standard Test Method for Measuring Deflections using a Portable Impulse Plate Load Test Device.” E2835-11, ASTM International, West Conshohocken, PA.

Automotive Industry Action Group (AIAG). (2010). Measurement System Analysis (4th ed). AIAG, Detroit, MI.

Cowles, M., and Davis, C. (1982). “ On the origins of the .05 level of statistical significance.” *American Psychologist*, 37(5), 553-558.

Dhawale, R. and Raut, D.N. (2013). ‘Evaluating measurement capabilities by gauge R&R using ANOVA for reliability.” *International Journal of Engineering Research and Application* 3(3), 726-730.

Fay, D.S., and Gerow K. (2013). “A biologist’s guide to statistical thinking and analysis.” WormBook, Available: http://www.wormbook.org/chapters/www_statisticalanalysi/statisticalanalysis.html (Last Accessed August 5, 2015).

Gertsman, B.B. (2006). “Introduction to Hypothesis Testing”, StatPrimer, Available: www.sjsu.edu/faculty/gertsman/StatPrimer/hyp-test.pdf (Last Accessed June 16, 2015).

Institute for Digital Research and Education (IDRE), (2015). “Faq: what is the coefficient of variation.” University of California at Los Angeles, (UCLA), Available: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/coefficient_of_variation.htm (Last Accssed July1, 2015)

Joubert, J.W. & Meintjes, S. (2015). Repeatability & reproducibility: implications of using GPS data for freight activity chains. Transportation Research Part B: Methodological, 76, 81-92. DOI: 10.1016/j.trb.2015.03.007

Mann, P.S. (2010). *Introductory Statistics, 7th Edition*. John Wiley and Sons, Inc, Hoboken, NJ.

Mazari M., Garcia, G., Garibay, J., Abdallah, I. and Nazarian. S. (2013). “Impact of modulus based device variability on quality control of compacted geomaterials using measurement system analysis.” *Proceedings of the Transportation Research Board 92nd Annual Meeting*, Washington D.C. 13p.

McLain, K. (2015), “Optimizing the Value of Soil Compaction Testing Quality Assurance and Control Using Stochastic Life Cycle Cost, Comparative and Statistical Analysis”, dissertation, ProQuest/UMI, Iowa State University.

Missouri State Highway Department, (MSHD). (1975). “Field evaluation of a direct transmission type nuclear moisture-density gauge.” *Missouri Cooperative Highway Research Program Final Report No. 74-2,* Missouri State Highway Department, Jefferson City, Missouri.

Montgomery, D.C., Runger, G.C., and Hubele, N.F. (2007). Engineering Statistics, Wiley and Sons

Nazzal, M. D., Abu-Farsakh, M.Y, Alshibli, K. and L. Mohammad L. (2007) “Evaluating the light falling weightdeflectometer device for in situ measurement of elastic modulus of pavement layers.” *Transportation Research Record: Journal of the Transportation Research Board*, No. 2016, Transportation Research Board of the National Academies, Washington, D.C., 13–22.

Nuzzo, R. (2014) Scientific method: Statistical errors, *Nature.com*, http://www.nature.com/news/scientific-method-statistical-errors-1.14700 (July 23, 2015)

Pandiripalli, B. (2010). “Repeatibility and reproducibility studies: a comparison of techniques”. M.S. Thesis,Wichita State University, Wichita, KS.

Seltman, H.J. (2015) “Chapter 7: one-way ANOVA.” *Experimental Design and Analysis*, Carnegie Mellon University, Pittsburgh, PA. 171-190.

Stamm, S. (2013). “ A comparison of gauge repeatability and reproducibility methods”.PhD Dissertation, Indiana State University, Terre Haute IN.

StatTools version 6.1.1, (2013), (computer software), Palisades Corporation, Ithaca, New York.

Wheeler, D. J. (2009). “An honest gauge R&R study.” Paper presented at the 2006 American Society for Quality/American Statistical Association Fall Technical Conference, Columbus, OH, http://www.spcpress.com/pdf/DJW189.pdf (April 5, 2015).

White, D.J., Vennapusa, P.K., Zhang, J. Geieslman, H., and Morris M. (2009). “Implementation of intelligent compaction performance based specifications in Minnesota.” Minnesota Department of Transportation Research Services Section, St. Paul MN.