NCAT Report 23-02

June 2023

National Center for Asphalt Technology
NCAT
at AUBURN UNIVERSITY

Evaluation of IDEAL-CT Testing Equipment

Final Report

Nathan Moore, Adam Taylor

Evaluation of IDEAL-CT Testing Equipment

Final Report

Southeastern Superpave Center Report

By

Nathan Moore, P.E., Principal Investigator
Assistant Research Engineer

Adam Taylor, P.E.
Assistant Research Engineer

National Center for Asphalt Technology at Auburn University

277 Technology Parkway

Auburn, AL 36830

June 2023

ACKNOWLEDGEMENTS

The authors would like to thank the representatives from ALDOT, FDOT, GDOT, KYTC, MDOT, NCDOT, SCDOT, and TDOT for their sponsorship, mixture donation, and feedback on this project. NCAT technicians Madison Eason and Tina Taylor are recognized for their meticulous attention to detail during specimen preparation and testing. Finally, the authors sincerely appreciate Ali Regimand from InstroTek, Inc., Clint Van Winkle and Finch Troxler from Troxler Electronic Laboratories, Todd Arnold from Pine Test Equipment, Inc., Andrew Cooper with James Cox & Sons, Inc., and Tripp Caldwell and Mahir Al-Nadaf with Humboldt Mfg. Co. for their support in this work. The authors gratefully acknowledge the members of the NCAT Application Steering Committee for their review of this technical report: Heather Hall, Elizabeth Pastuszka, Katherine Erwin, and Stacey Diefenderfer.

DISCLAIMER

The contents of this report reflect the views of the authors who are responsible for the accuracy of the data and analysis presented herein. The contents do not necessarily reflect the official views or policies of the National Center for Asphalt Technology or Auburn University. This report does not constitute a standard, specification, or regulation. Comments contained in this paper related to specific testing equipment and materials should not be considered an endorsement of any commercial product or service; no such endorsement is intended or implied**.**

ABSTRACT

The IDEAL-CT has become a popular cracking test in the asphalt industry. Many sources of variability in the IDEAL-CT testing results have been studied but a comprehensive assessment of the testing equipment remained lacking. NCAT evaluated six different IDEAL-CT testing devices to assess if they complied with the ASTM D8225-19 IDEAL-CT specification and to see how the results from different devices compared with each other. A total of 328 tests were conducted from seven unique asphalt mixes. Each specimen was prepared with meticulous attention to detail to minimize the variability from specimen preparation, thus magnifying the variability from the devices themselves. With the inherent variability of the IDEAL-CT, it is impossible to expect two devices to produce results with perfect agreement. Thus, the Two One-Sided Tests (TOST) equivalence test was conducted to determine whether the devices could be considered equivalent. The TOST is a procedure designed to compare testing processes that possess measurable variability and accounts for this variability during the comparison. This procedure was used to compare IDEAL-CT testing devices in this study.

Four of the six devices consistently operated at deformation rates outside of the required rate of 50 ± 2.0 mm/min. However, all six devices yielded a measured rate of 51.0 ± 2.0 mm/min. The differences in the devices' rates had no effect on the testing results. Equivalence between the devices was accepted for all devices except one. In this case, the specific manufacturer discovered an issue with their device and made appropriate changes to resolve the issue.

Finally, as a result of this work, all of the device manufacturers made changes to their equipment to either bring them in compliance with the specification or to make their products more user-friendly. These changes have since been updated in the devices currently in use in the industry. Thus, as a result of this work, the available equipment has been improved.

TABLE OF CONTENTS

INTRODUCTION

The IDEAL-CT (InDirect tEnsile Asphalt Cracking Test) has become one of the more popular cracking tests in the asphalt testing community since it was developed in 2017 (Zhou, et al., 2018). The test was developed considering seven desirable features identified in NCHRP project 9-57, which included simplicity, relatively low-cost testing equipment, and a repeatability coefficient of variation (COV) less than 25% (Zhou, et al., 2016).

The ease of use and cost of equipment are two reasons many agencies have either already moved towards adopting it as a required mix design tool or are considering doing so (Yin and West, 2020). Not long after the introduction of the IDEAL-CT, contractors and agencies began using it in their labs. As more users grew familiar with the test, questions regarding testing equipment soon followed. Many people asked, "Can I use my current load press?" or "Is there a difference between my machine and the state's machine?" Most users assumed that all devices were the same if they could load an IDEAL-CT specimen at the specified rate of 50 mm/min.

NCAT researchers began using extra specimens from ongoing projects and testing them on multiple NCAT lab devices to better understand the issue. A small-scale experiment was conducted using three mixes and two loading devices to generate more data to help answer the device-to-device variability and equality question. These results were combined with five other mixes that had been tested on both machines. CT_{Index} values for these eight mixes are shown in Figure 1. The CT_{Index} was higher for seven out of eight of the mixes and the fracture energy was higher for all eight mixes on one of the devices. These initial results cast doubt on the idea that all devices were equal and could be trusted to provide the same results. If multiple devices could not be expected to produce consistently similar results for the same mix, it would be essential for users to be aware of this issue. This necessitated a framework to be developed to assist users in identifying non-equivalence among devices. Thus, a more extensive study was planned to investigate the issue further.

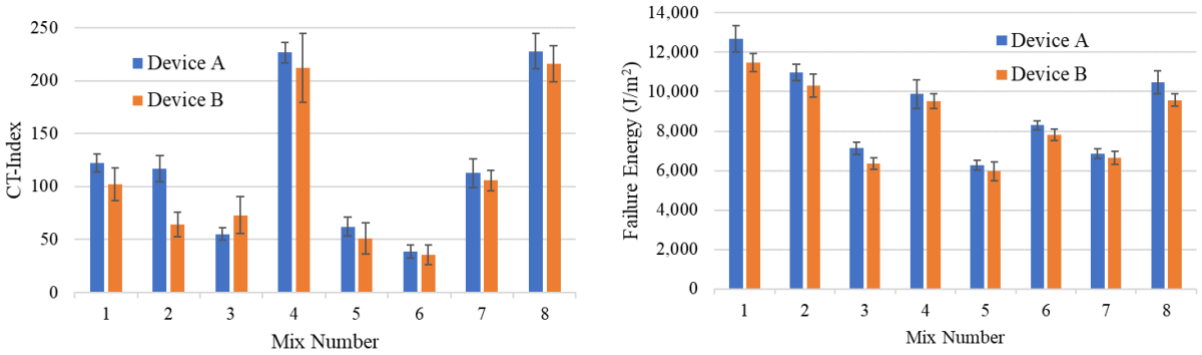

Figure 1: Initial IDEAL-CT device comparison results from NCAT

Several inter-laboratory studies (ILS) and ruggedness tests have been conducted to ensure repeatability and sensitivity for this test (Zhou et al., 2018, Taylor et al., 2022, Diefenderfer et al., 2020). A two-phase ILS for the IDEAL-CT was conducted by NCAT in 2019 using plant-mix

samples. Phase I required each participating lab to prepare, compact, and test the specimens. In contrast, in Phase II, all specimens were prepared and compacted by NCAT prior to sending to the participating labs for testing. The overall within-lab COV from both phases were remarkably similar, slightly less than 20%, indicating that the test is repeatable within a single lab. Other ILS's have reported similar repeatability values (Diefenderfer, et al., 2020). However, the between-lab COVs for the two phases of the NCAT ILS were dramatically different. The between-lab (i.e., the reproducibility) COV% for Phase I was 35%, but, the Phase II reproducibility COV dropped to approximately 20% (Taylor, et al., 2022).

Suppose the test repeatability is the same regardless of who makes the specimens and reproducibility can be dramatically improved by having all the specimens created by a single lab. In that case, the question remains: "How much of the remaining variability in CT_{Index} comparisons can be attributed to differences in the devices?" Considering the growing popularity of the IDEAL-CT test combined with the numerous options for testing devices, it was prudent to investigate sources of variability in the test that could create undesirable consequences after the test becomes entrenched in many agencies' specifications.

OBJECTIVES

The three study objectives were:

1. Assess whether testing devices meet the current ASTM specification (D8225-19).
2. Determine if the devices in this study produce equivalent results.
3. Provide guidance for the approval of testing devices for IDEAL-CT testing.

DEVICES, MATERIALS, AND METHODS

This section details the devices that were evaluated and the asphalt mixtures used in the process. It also outlines the specimen preparation methodology and data analysis techniques.

Devices Evaluated

This study assessed six commonly used IDEAL-CT devices as shown in Figure 2. All devices were within calibration at the time of testing or were loaned/donated with the factory calibration records for the study. Each device also had the most recent available firmware or software at the time of testing (May – July 2020). Four of the load frames were screw-driven loading devices and one (Troxler IDEAL Plus, formerly TestQuip) used servo-hydraulic loading. The InstroTek Smart-Jig was used in conjunction with the Pine 850T press. The Smart-Jig is not a load frame and can only record load and displacement data from an external machine. For this study, the Pine 850T device and the Smart-Jig recorded data independently from the same set of specimens. At the time of this study, the InstroTek load frame was named the Auto-SCB but has since been renamed the SmartLoader. This report will refer to it by its former name.

Figure 2: Devices used in study (from left to right): Top row: Cox & Sons 9500, Humboldt 5125, InstroTek Auto-SCB. Bottom Row: InstroTek Smart-Jig, Pine 850T (with InstroTek Smart-Jig), and Troxler IDEAL Plus

Before comparisons of device-to-device results could be addressed, each device was assessed for compliance with the specification. The devices from Troxler, Humboldt, Cox & Sons, and InstroTek were all developed specifically for the IDEAL-CT test. The Pine 850T press has the same mechanics as the older load presses that use the paper data recording system, but was upgraded to a digital recording system to meet the ASTM D8225 specification device requirements listed in Table 1. ASTM D8225-19, at the time of testing, required the following elements related to the devices.

Table 1: IDEAL-CT Device Requirements (per ASTM D8225-19)

Device Feature	Requirement
Axial Loading Device - Capacity	Min. 25 kN (5,620 lbs)
Axial Loading Device – Deformation Rate	Constant 50 ± 2.0 mm/min
Load Cell – Resolution	10 N
Internal/External Displacement Measuring Device – Resolution	±0.01 mm
Data Acquisition System - Frequency	Min. 40 Hz.

Asphalt Mixtures Tested

Seven mixes were tested for device-to-device comparisons. Although the contents and comparison between the mixes were not the focus of the study, each mix was a production mix sampled from paving projects across the Southeastern U.S. Several of the mixes were selected based on CT_{Index} values reported by the supplying agency or contractor to provide a range of CT_{Index} values from approximately 30 to 150. However, when the mixes were actually tested in the NCAT lab, the actual range of CT_{Index} results was only 40 – 80. Although this was unfortunate, these values are still representative of mixes produced in the Southeastern U.S. Information regarding typical mix parameters are provided in Table 2. As previously stated, the mixes themselves were not the focus of the study; they were merely used to assess device variability.

Table 2: Mixtures Tested in Device Evaluation

Mix ID	NMAS	AC%	RAP%	PG-Grade	Expected CT_{Index}	Actual CT_{Index} (Approx.)
A	12.5	5.3	31%	67-22	30	50
B	9.5	6.2	35%	67-22	57	40
C	12.5	5.4	30%	64-22	65	50
D	9.5	5.7	35%	67-22	65	50
E	9.5	6.1	35%	67-22	150	70
F	9.5	6.2	30%	70-22	94	60
G	9.5	5.5	30%	64-22	100	80

Specimen Preparation

Extraordinary attention to detail was given to achieve consistent specimen preparation to minimize variability due to specimen preparation, which could mask the variability source of interest in this study, which were the devices themselves. For each of the seven mixtures, 60 to 70 IDEAL-CT specimens were prepared by the same technician using the same scale, oven, location and time in the oven, gyratory compactor, and two gyratory molds, conditioning chamber, and conditioning time. The 48 specimens with air voids closest to the average air void content for all specimens within that mix were selected and randomly assigned to six groups of eight replicates. All specimens sat at ambient laboratory temperature for two weeks before testing. After the two week period elapsed, testing was completed within a 24-hour window for each mix.

The difference from the average air voids of all the specimens for each mix was used as the basis for specimen selection instead of the typical value of 7.0% to minimize the spread in air void contents among the specimens. It has been reported elsewhere that air void content is a source of variability in this test (Zhou, et al., 2018, Chen, 2020), which is why the focus was on limiting the spread instead of meeting a specific target. The average and range of air voids for each mix are listed in Table 3. Note that some specimens from Mixes B, C, and G had air void contents outside of the range specified in the ASTM D8225 standard. However, because the purpose of the study was to evaluate differences between test devices and not mixes, it was

more important to have the smallest range of air voids possible instead of having specimens at 7.0 ± 0.5.% air void contents. The largest spread of air voids seen in any of the seven mixtures was ± 0.4%, which was lower than the ASTM threshold of ± 0.5%.

Table 3: Summary of individual specimen densities

Mix	Avg. Air Voids, %	Air Void Range, %
A	7.3	7.1 – 7.5
B	6.6	6.4 – 6.8
C	6.8	6.4 – 7.2
D	7.0	6.8 – 7.2
E	6.9	6.7 – 7.1
F	7.2	7.0 – 7.3
G	7.5	7.1 – 7.9

Analysis Methods

All testing data was processed through the software provided by the manufacturer. In other words, no data cleanup was applied, or additional calculations performed through a separate template or calculator. The only data post-processing conducted was applying ASTM E178-16 as a standard outlier evaluation procedure. This was performed on every set to eliminate any outliers discovered at a 90% two-tailed confidence level, per NCAT's standard practice. In total, 328 tests were performed on 280 specimens (the Smart-Jig and the Pine 850T press provide two unique measurements on the same specimens). Only seven test results were flagged as outliers and discarded from the analysis.

Each device, except the Pine 850T press, had an external displacement measuring device that measured the testing speed during the test. The Pine 850T calculated the position of the platen based on the motor speed and time and then corrected the displacement measurement for machine deflection under high loads. The correction methodology is detailed in Pine Technical Bulletin #041618 (Pine, 2018). Internal testing at NCAT demonstrated that this calculated correction was less accurate than the external LVDT on the InstroTek Smart-Jig. The average deformation rate from Pine was 54.6 mm/min, which was 2.1 mm/min faster than the rate measured by the Smart-Jig. Thus, the testing rate calculated by Pine was replaced with the rate measured with the Smart-Jig to provide a more accurate estimate of the Pine 850T testing rate. The deformation rates for each device were analyzed and compared to the specification. Finally, the sampling frequencies of the devices were assessed by counting the number of data points recorded per second.

Determining whether two unique devices can be trusted to produce similar results does not require that the devices produce equal results. Due to differences in these devices, it is expected that for a given sample of asphalt mix one device will produce a CT_{Index} that is different from another device. The magnitude of the difference compared to the random variability of the test is what is important. For example, if two devices tested an asphalt mix and the CT_{Index} results were 70 and 72, practitioners would consider these results practically equal. Furthermore, if two devices are equivalent, each of them will produce a larger result in head-to-

head testing approximately 50% of the time. This idea that devices can be considered equal except for very small and irrelevant differences is referred to as "equivalence" (Wellek, 2010). For this study, the Two One-Sided t-tests (TOST) procedure was used to evaluate device equivalence. More information about equivalence testing, specifically regarding how it was used in this study, is published in Moore, et al., 2022.

Equivalence testing requires an estimate of the variability of the test method. From other variability studies of the IDEAL-CT (Taylor, et al., 2022, Diefenderfer, et al., 2020), the within-lab COV of the test has been shown to be approximately 20%. Thus, test results produced by the same operator should be considered acceptable if their standard deviation is within 20% of the mean results. Tests with greater precision will have smaller ranges for values considered equivalent. IDEAL-CT devices should not be expected to agree with greater precision than the test is capable of consistently producing. To test for equivalence using the TOST method, the user must select an equivalence limit "E" where estimates of the difference between two devices that exceed this limit are considered non-equivalent. When the 90% confidence interval of the estimate of the difference between devices is less than the equivalence limit, the inherent variability from the IDEAL-CT test overshadows the difference estimate. In these cases, the difference between devices is smaller than the accepted variability of the test. Therefore, devices with a difference less than the equivalence limit cannot be considered non-equivalent. The equivalence limit for this study was set at 20% of the average CT_{Index} results for all mixes combined because 20% COV is the assumed single-operator acceptable repeatability. Thus, if the entire 90% confidence interval for the difference estimate between two devices was less than 20% of the average CT_{Index}, the devices were considered equivalent. This methodology is covered in greater detail in Moore, et al., 2022 and is not repeated in this report.

RESULTS AND DISCUSSION

The results of this study are presented further in two sections. The first section focuses on the devices' compliance with the specification, specifically the required deformation rate. The second section presents the results from the IDEAL-CT testing for the devices.

Compliance with ASTM D8225-19

Each device met the requirements for the loading capacity, load cell resolution, and displacement measurement device resolution. The only two requirements that had non-conformity among the six devices included in this study were the deformation rate of the loading device and the frequency of the data acquisition system.

Data Acquisition System Frequency

After completing the testing, it was noticed that the Humboldt 5125 and the Cox & Sons 9500 devices had inconsistent data sampling frequencies. ASTM D8225-19 requires the sampling frequency to be a minimum of 40 data points per second, or 40 Hz. The recorders of these two devices would capture between 35 and 41 data points during the first second of the test and then would average approximately 39.7 Hz for the duration of the test. This information was

shared with the manufacturers and appropriate modifications were made immediately. It was not expected that this issue would have a noticeable effect on the final results.

Axial Loading Device Deformation Rate

ASTM D8225-19 states "The loading apparatus… shall be capable of maintaining a constant deformation rate of 50 ± 2.0 mm/min…" There is a common misconception that deformation is the same as displacement. In asphalt testing devices, displacement is the distance traveled by the loading strips or the platen. Deformation is the distance that a specimen actually deforms during loading. The difference becomes important when loads are high enough to bend or compress the test frame. Some energy that is supposed to be transferred to the specimen during loading is lost to the machine frame as it also deforms.

The specification requires maintaining a constant deformation rate between 48 and 52 mm/min. Practically this requires a closed-loop feedback system on the device, which is even suggested, but not required, in the ASTM specification. Without a closed-loop feedback system on the loading rate, the devices tend to operate below the target rate during loading until the peak load and then overshoot the target rate as the specimen breaks and the load decreases. A helpful analogy of this behavior is to compare this to how a motor vehicle operates driving up and down a hill. Suppose a driver is targeting a specific speed by keeping the gas pedal at a constant position when the vehicle encounters a hill. In that case, its speed will decrease because it is meeting more resistance than it did on a flat section of roadway. When the vehicle reaches the top of the hill, it will resume the target speed until it begins to descend the hill. At this point, the vehicle will exceed the target speed because there is less resistance than there was on the flat section. This is the typical behavior of screw-driven loading devices. During increasing loads, the device is slower than the target rate, and at the peak load, the loading devices operate near the target rate. As the load decreases past the peak load, the device's speed increases. A closed-loop feedback system prevents this behavior, much like cruise control in a vehicle allows a car to maintain a set speed by varying the power input depending on the grade of the roadway.

Figure 3 shows a typical load vs. deformation curve for the IDEAL-CT test overlaid with the deformation rate of the devices during the test to demonstrate how each device in this study behaved during loading. Deformation rates were calculated using a moving average over an interval of 0.3 seconds. As discussed previously, the Pine 850T does not actually measure displacement; this device calculates its position based on the motor speed and time and corrects the displacement measurement to estimate specimen deformation. Thus, the line representing the InstroTek Smart-Jig is a better estimation of the behavior of the Pine 850T device because it contains an external LVDT that measures the true position of the Pine device. All but two of the devices spent a significant portion (>50%) of the testing time outside of the specified rate range. The two devices that complied with the specified deformation rate through most of the loading sequence were the Troxler device, which has a closed-loop feedback system and was the only servo-hydraulic device in the study, and the InstroTek Auto-SCB, which also had a form of a closed-loop feedback system to regulate speed.

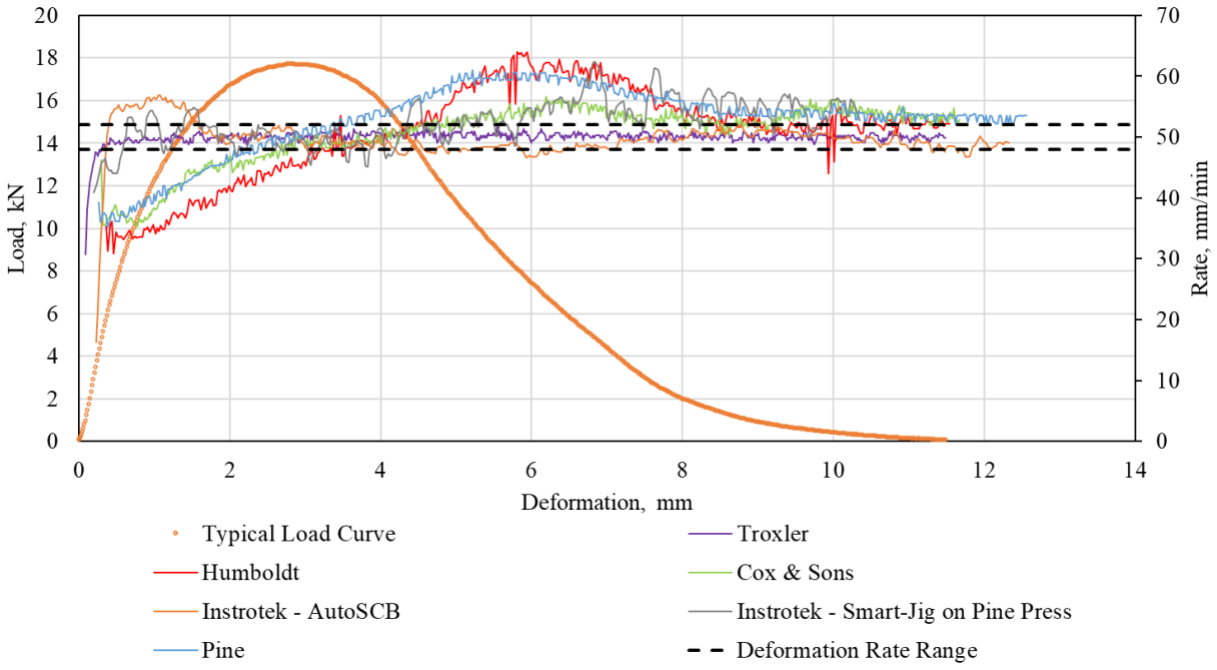

Figure 3: Example of Load vs. Displacement Curves for Each Device

Although displaying the instantaneous rate for a device during the course of the test provides a clear visual description of how the device is behaving, most manufacturers simply report the average speed over the entirety of the test. The averages and standard deviations of the speeds of the six devices in this study are shown in Table 4. Again, note that the speed listed for Pine is not as accurate as the speed listed for the Smart-Jig, which was used on the Pine 850T press. Only three devices had averages in the required specification range. 100% of the replicates for the Auto-SCB and the Troxler devices met the specification along with 67% of the replicates for the Cox & Sons device.

Table 4: Device Speed Summary

Device	Average Speed, mm/min	SD Speed, mm/min	Number of Tests	% of Tests in ASTM D8225-19 Range
Auto-SCB	49.5	0.1	55	100%
Cox & Sons	51.8	0.3	55	67%
HM-5125	52.6	0.3	55	0%
Pine 850T	54.6	0.3	54	0%
Smart-Jig*	52.5	0.2	48	4%
Troxler	50.1	0.0	54	100%

* Smart-Jig in Pine 850T load frame

Figure 4 shows a histogram of the speeds measured in this study, excluding the Pine 850T results. Interestingly, although 55% of the data are above the maximum speed allowed in ASTM D8225-19, all 267 speed measurements were within a 4 mm/min range. The minimum speed

was 49.3 mm/min, and the maximum measured speed was 53.1 mm/min. Thus, all devices operated within a range of ± 2 mm/min but not at the specified target. The next section presents the IDEAL-CT test results and analyzes the effects of speed on the final CT_{Index}. If the results are not affected by the differences in the speeds, then a specification change to allow all these devices could be warranted.

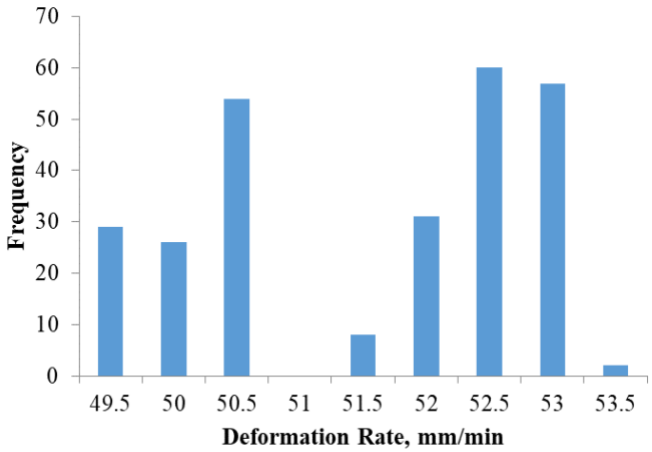

Figure 4: Histogram of Device Deformation Rates

IDEAL-CT Results

The results for the seven mixes across the six devices are shown in Figure 5. The error bars represent one standard deviation of the CT_{Index}. A consistent y-axis across all mixes was intentionally avoided. This technique is commonly used when the purpose of the study is to compare test results across different mixes. However, since each mix represents a separate analysis and the purpose of the study is to compare the devices, the y-axis was set to highlight the differences for each device for each mix individually.

A table of summary statistics for the mixes is provided in

Table 5. The devices for this portion of the analysis were deliberately concealed, as agreed upon by the researchers, sponsors, and equipment manufacturers. The purpose of presenting these data was to identify if any devices produced consistently higher or lower results than the rest. The device numbering system utilized in this section was randomly generated. Examination of Figure 5 shows that Device #5 consistently produced the lowest CT_{Index} across five of the six mixes; an operator error prevented Mix G data on Device #5 from being collected. However, it is clear that Device #5 is consistently different from the other five devices. Table 5 summarizes the statistics for the six mixtures for all devices combined. Table 6 summarizes the

average COV for the six devices for all mixes combined. Each device had repeatability consistent with results from previous ILSs and Round Robin studies.

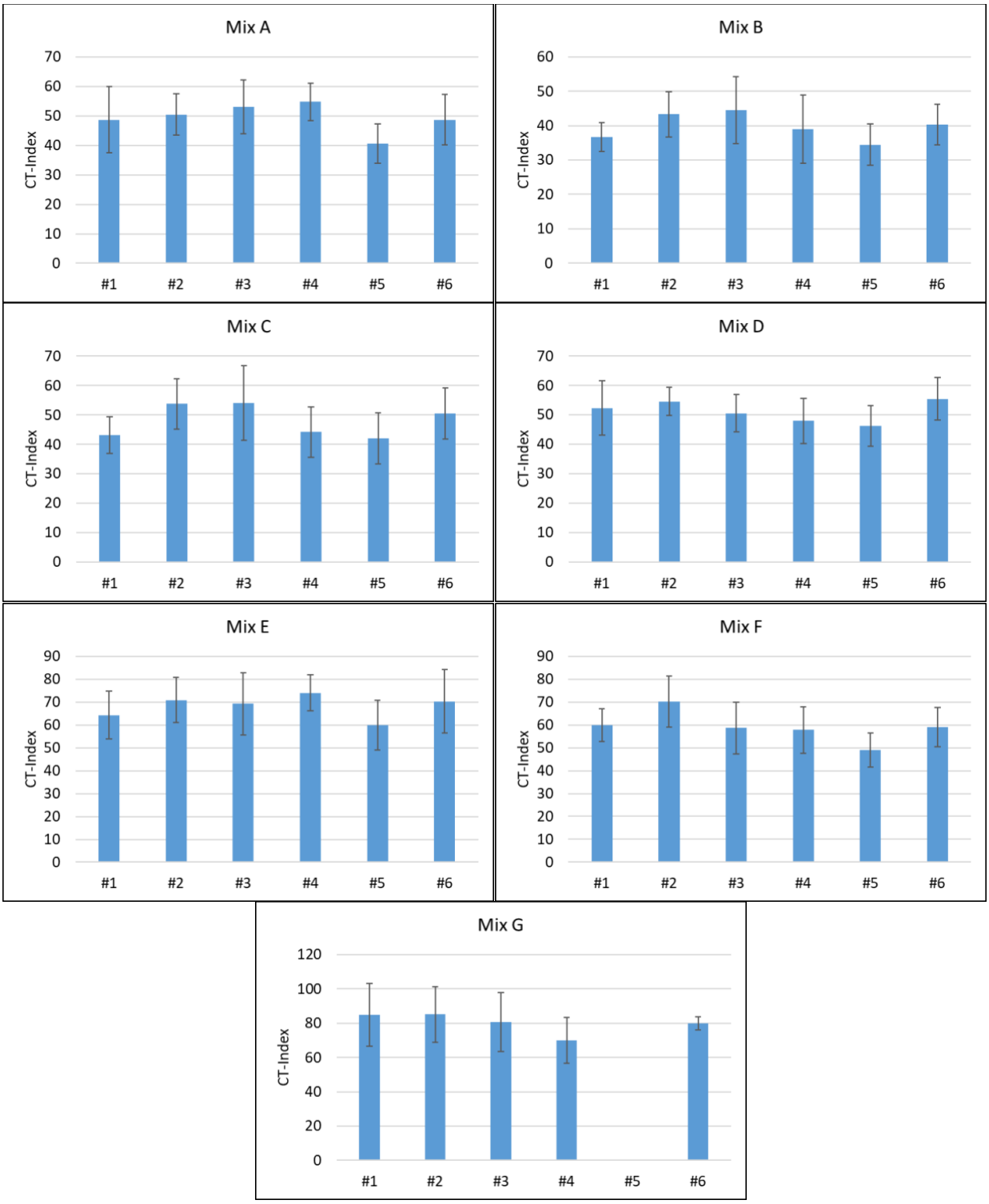

Figure 5: CT_{Index} results for each mix (A-G) and device (1-6). ***Error bars represent one standard deviation***

Table 5: Summary of Results for Mixes A-G

Mix	Avg. CT_{Index}	Std. Dev. CT_{Index}	COV, %
A	49.4	9.1	18%
B	39.8	7.9	20%
C	47.9	9.9	21%
D	51.2	7.8	15%
E	68.0	11.7	17%
F	59.1	10.9	18%
G	80.0	15.1	19%

Table 6: Summary of Device Repeatability

Device #	Avg COV, %
1	16.6%
2	14.7%
3	19.2%
4	17.1%
5	17.1%
6	14.5%

As shown previously, the deformation rates of some devices were not within the 50 ± 2.0 mm/min range. However, the devices had deformation rates within 51.1 ± 2.0 mm/min. Figure 4 shows how the rates were split into two groups that centered around approximately 50 mm/min and 52.5 mm/min. A two-sample t-test conducted to compare the mean CT_{Index} of these two groups resulted in a p-value of 0.882, indicating that the mean CT_{Index} results of the two groups were not statistically different. Furthermore, linear regressions were conducted on the CT_{Index} results and the deformation rates for each specimen for each mix and all mixes combined. The largest R^2 result from any of the mixes was 0.248 (Mix E) while all other mixes had R^2 values less than 0.01. The R^2 value for all specimens combined was 0.001. These results indicate that the deformation rates did not explain the testing variability and did not significantly affect the final CT_{Index} results for the mixes. This is almost certainly because the devices were operating within a range consistent with the deformation rate tolerance (± 2 mm/min) in ASTM D8225-19 that was based on the ruggedness testing from the original IDEAL-CT study (Zhou, et al., 2018).

Finally, recall that the equivalence limit (shown as Δ in Figure 6) was set at 20% of the average of all CT_{Index} results in this study. The global average of all specimens was 55, so the equivalence limit was ± 11 CT_{Index} units. Two devices were considered to provide equivalent results when the entire 90% confidence interval for the estimated difference between two devices was within ± 11 CTIndex units. In total, fifteen possible pairwise comparisons were made between the average CT_{Index} measurements from the six devices. The results from the TOST test are shown in Figure 6. The x- and y-axes are the CT_{Index} measurements of two different devices. The points in the figure represent the average CT_{Index} for Device x (on the x-axis) vs. the average CT_{Index} for Device y (on the y-axis). The lines protruding in both directions from the point represent the 90% confidence interval for the difference between the two devices. The blue region represents the values that are within the equivalence limits. If the point estimate for the pairwise comparison and the entire 90% confidence interval are within the shaded blue region, those two devices are considered to provide equivalent results.

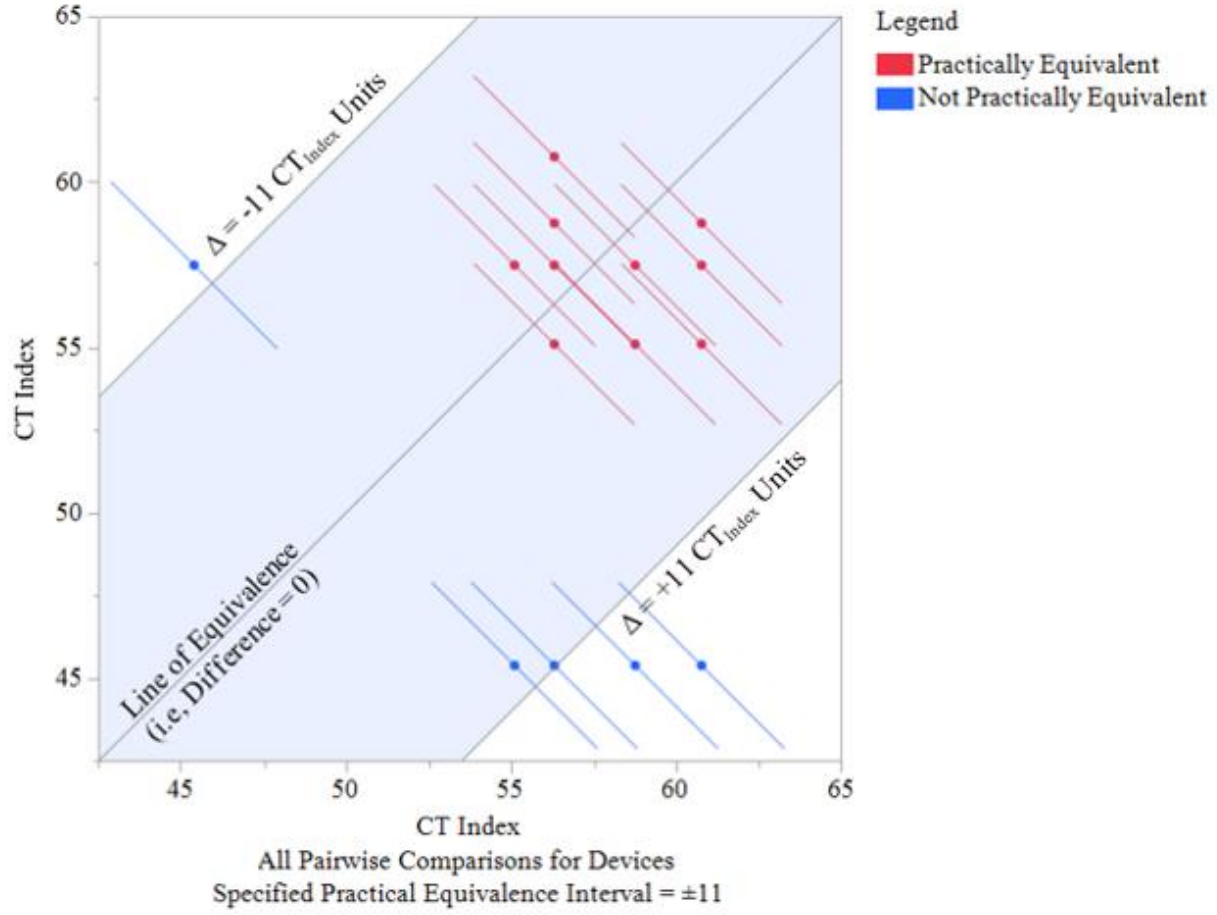

Figure 6: TOST Equivalence Results for All Device Comparisons

Five of the fifteen comparisons produced non-equivalent results. These are the blue points and lines in Figure 6. They all include a comparison with Device #5. All other comparisons produced equivalent results. Thus, for this study, it is concluded that despite the differences in

deformation rates, all of the IDEAL CT devices provide equivalent results, with the exception of Device #5. All other device comparisons had average differences less than 6 CT_{Index} units. It should be noted that there were instances of double-digit differences between pairwise comparisons that did not include Device #5. However, these instances were not consistent, and the effect of the large differences was diminished by combining all the difference estimates together.

The results from each specific device were presented to the manufacturer representatives, and a preliminary analysis was shared. For all other devices except Device #5, other issues of non-conformity with the ASTM specification and user-friendliness of software were discussed. When the results from Device #5 were shared with that manufacturer, they recognized an error in their data recording, and the issue was resolved. They corrected the issue for all devices with the same issue currently in use. NCAT re-tested the machine after the correction and confirmed that the issue had been resolved and the results were much improved. These results are not shown in this report.

One drawback of the TOST procedure is the relatively large number of samples required compared to what is typically used in the industry to determine whether two devices provide equivalent results. A statistical power analysis was performed to determine how many replicates would be necessary to perform the TOST correctly. If good within-lab repeatability (COV ≈ 20-25%) is achieved, then 40 - 50 replicates for each device will be enough to determine equivalence with the TOST procedure. The number is highly dependent on the variability of the IDEAL-CT tests and the difference between the two devices being analyzed. This is a conservative estimate of the required number of replicates, and improved repeatability will reduce this number. However, if the number of replicates tested is too low, the TOST will result in non-equivalence simply because it is designed to assume devices are non-equivalent unless the data prove otherwise. The same number of replicates from each mix should be tested for this procedure. For example, if four replicates from a mix are tested on one device, the same number should be tested on the other. These data can be collected relatively easily over the course of a month by collecting testing results over time into a database.

IDEAL-CT users, especially contractors, should keep a log of each specimen tested along with summarized testing information for each mix (averages, standard deviation, number of replicates, etc.). This will help expedite the process if a discrepancy between two devices occurs. A more detailed description of how the TOST can be implemented to assess device equivalence in practice is presented in Moore et al., 2022. The TOST procedure can be performed using most statistical analysis software such as Minitab, SPSS, JMP, etc. Finally, ASTM E2935-20e1: *Standard Practice for Conducting Equivalence Tests for Comparing Testing Processes* can also be used to perform the TOST procedure. A summarized method for determining whether two devices are equivalent is provided below:

1) Identify the devices that seem to be producing different IDEAL-CT results.
    a. Are the devices running at appropriate speeds?
    b. Are the devices recording the same units and/or calculating CT_{Index} correctly?

    c. Are the specimens being tested under the same conditions?
2) Compile the recent testing results from identical mixes (i.e., split samples) that were tested by both mixes around the same time.
    a. The number of replicates from each mix should be approximately equal.
    b. If the within-lab repeatability COV estimates of the two devices are both <25%, 50 specimens will be sufficient to analyze whether the devices are equivalent. If there is poor repeatability, investigate differences in specimen preparation, handling, and testing between the two labs.
    c. If more replicates are needed, it is highly recommended that more be made in the same lab and then randomly distributed between the two devices.
3) Perform the TOST, as listed in ASTM E2935-20e1, and set the equivalence limit to 20% of the overall average of all the combined specimens.
    a. Note: if the number of replicates tested are too few, the TOST will likely result in non-equivalence. It is critical that enough specimens are tested for this procedure to be accurate.
4) Accept or reject device equivalence based on the results from the TOST procedure.

CONCLUSIONS AND RECOMMENDATIONS

A total of 328 IDEAL-CT tests were conducted on six different devices to assess 1) how well each device met the ASTM D8225-19 specification and 2) how the results compared between different devices. The following conclusions are made:

- The data sampling frequencies observed in this study did not appear to have affected the final results, even when they did not meet the ASTM D8225 specification requirements.

- It was determined that four of the devices consistently operated at deformation rates outside of the range specified in ASTM D8225-19. However, all rates measured with external measurement devices were within 51.1 ± 2.0 mm/min. All the devices met a ± 2.0 mm/min tolerance but the specific range was not within the allowable 48 to 52 mm/min in the current ASTM D8225 specification. It is recommended that ASTM change the D8225 specification to accept devices that produce average deformation rates up to 53.0 mm/min. This recommendation is based on the fact that the device deformation rates had little to no effect on the results in this study within the range of deformation rates encountered in this project.

- The comparison of six devices indicated that most provided equivalent CT_{Index} results. One device produced CT_{Index} results that were not equivalent to all other devices using the TOST equivalence procedure. The issue causing the lack of equivalence for that device was corrected by the manufacturer after the results of the study were presented to the manufacturer. The device was reevaluated and the results improved after the manufacturers corrected the issue. The CT_{Index} results for each of the other five devices were determined to be equivalent based on the TOST procedure. This study only used

one device from each manufacturer. It is possible that the results could be different on other similar models.

- A method for determining equivalence between any two devices was proposed. In summary, an equal number of specimens from an equal number of mixes should be tested for both devices in question. Provided the within-lab repeatability is less than COV=25%, 50 specimens for each device are a conservative estimate of the total number of samples required to conduct the TOST equivalence test. This test can be conducted using most statistical analysis software. A detailed test procedure is found in ASTM E2935-20e1: *Standard Practice for Conducting Equivalence Tests for Comparing Testing Processes.*

- Valuable feedback was provided to the manufacturers of the IDEAL-CT testing devices from this study. The manufacturers implemented much of this feedback, resulting in improved quality IDEAL-CT testing equipment available to users on the market.

CHANGES IMPLEMENTED BY MANUFACTURERS

The results of this study were shown to the manufacturers, and each was provided with feedback specific to their devices. In several cases, the only feedback suggested was minor changes regarding the user-friendliness of software or testing reports. However, in the case of nonconformity issues with the ASTM D8225-19 specification, the manufacturers made the necessary changes to correct issues as best as possible. The manufacturers' comments regarding specific equipment changes are included in the Appendix of this report. The issue with the deformation rates falling outside the ASTM D8225-19 target range was reported here to be inconsequential, as it did not appear to have any effect on the testing results because, although different, they were still within a ± 2.0 mm/min tolerance.

REFERENCES

ASTM. ASTM Standard E2935-20e1: Standard practice for conducting equivalence tests for comparing testing processes. ASTM International. West Conshohocken, PA. 2020

Chen, C., Validation of Laboratory Cracking Tests for Field Top-Down Cracking Performance. Ph.D. Dissertation, Auburn University, 2020

Diefenderfer, S.D., Boz, I., Habbouche, J., and Bilgic, Y.K. Technical Memorandum: Round Robin Testing Program for the Indirect Tensile Cracking Test at Intermediate Temperature - Phase I. Virginia Transportation Research Council, Charlottesville, 2020, 9p.

Moore, N., Steger, R., Bowers, B., & Taylor, A. Investigation of IDEAL-CT Device Equivalence: Are All Devices Equal? Transportation Research Record Vol. 2676(5), pp. 1-12. 2022. https://doi.org/10.1177/03611981221091551

Pine. Technical Bulletin 041618. Pine Test Equipment. Grove City, PA. April 16, 2018

Taylor, A., Moore, J., and Moore, N., NCAT Performance Testing Round Robin, NCAT Report 22-01. National Center for Asphalt Technology, Auburn, AL. 2022.

Wellek, S., Testing Statistical Hypotheses of Equivalence and Noninferiority. Chapman and Hall/CRC, Boca Raton, FL, 2010

Yin, F. and West, R. Balanced Mix Design Resource Guide – IS-143. National Asphalt Pavement Association. Greenbelt, MD, 2020

Zhou, F., Im, S., Sun, L., & Scullion, T. Development of an IDEAL cracking test for asphalt mix design and QC/QA. Road Materials and Pavement Design, 2018. 18(sup4), 405-427.

Zhou, F., Newcomb, D., Gurganus, C., Banihashemrad, S., Park, E.S., Sakhaeifar, M., and Lytton, R., Experimental design for field validation of laboratory tests to assess cracking resistance of asphalt mixtures (Final Report). National Cooperative Highway Research Program Project 9-57. Washington, D.C. Transportation Research Board of the National Academies. 2016

APPENDIX

The device manufacturers reviewed an initial draft of this report. They were given the opportunity to provide further details regarding specific changes they made to their testing equipment as a result of this study. These responses are included below in alphabetical order of the manufacturers.

Humboldt Mfg. Co.

"Initially the system was not using a control algorithm and was only using a static speed which did not account for the machine stretch. Since working with [NCAT] we have seen the need for this, added feedback control, and… have further refined it."

Instrotek, Inc.

"We increased the load cell sensitivity in the lower load region."

James Cox & Sons, Inc.

No feedback provided.

Pine Testing Equipment

"The Pine 850T referred to in this report ran at a deformation rate higher than allowed by the IDEAL-CT test method, ASTM D8225-19. Pine has modified the drive motor gear ratio to run at a slower deformation rate. The algorithm for determining displacement was not changed."

"NCAT tested specimens with [this] modified Pine 850T using the same asphalt mixtures used in this report. The average specimen deformation rate for these tests was 51.1 mm/min and all the data were within 50 ± 2 mm/min. Furthermore, Nathan Moore at NCAT conducted the same statistical analysis of IDEAL-CT test results as presented in this report, TOST, and concluded that there was [equivalence] between the IDEAL-CT generated by the modified Pine 850T and the results generated by other machines in the study."

Troxler Electronic Laboratories

No feedback provided.