CAPRI

CONSORTIUM FOR ASPHALT PAVEMENT RESEARCH AND IMPLEMENTATION

Evaluating the Impact of Specimen Fabrication Variabilities on Balanced Mix Design Performance Testing

capriasphalt.us

Evaluating the Impact of Specimen Fabrication Variabilities on Balanced Mix Design Performance Testing – Expanded Study

By

Aaron Leavitt, M.S., PE
Research Engineer, Program Manager
Federal Highway Administration
Turner Fairbank Highway Research Center, McLean, Virginia

S. Farhad Abdollahi, Ph.D.
Research Engineer,
Turner Fairbank Highway Research Center, McLean, Virginia

Behnam Jahangiri, Ph.D., PE
Research Engineer, Lab Manager
Turner Fairbank Highway Research Center, McLean, Virginia

May 2026

DISCLAIMER

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the use of the information contained in this document. Non-Binding Contents—Except for the statutes and regulations cited, the contents of this document do not have the force and effect of law and are not meant to bind the States or the public in any way. This document is intended only to provide information regarding existing requirements under the law or agency policies. Quality Assurance Statement—The Federal Highway Administration (FHWA) provides high-quality information to serve Government, industry, and the public in a manner that promotes public understanding. Standards and policies are used to ensure and maximize the quality, objectivity, utility, and integrity of its information. FHWA periodically reviews quality issues and adjusts its programs and processes to ensure continuous quality improvement. Disclaimer for Product Names and Manufacturers—The U.S. Government does not endorse products or manufacturers. Trademarks or manufacturers' names appear in this document only because they are considered essential to the objective of the document. They are included for informational purposes only and are not intended to reflect a preference, approval, or endorsement of any one product or entity.

Furthermore, the contents do not necessarily reflect the official views or policies of the National Center for Asphalt Technology or Auburn University.

ACKNOWLEDGEMENTS

The author would like to thank the Consortium for Asphalt Pavement Research and Implementation (CAPRI) for sponsoring this important research. Special thanks to members of the project oversight panel and participating laboratories that made the project possible.

Table of Contents

List of Figures

List of Tables

1 Background

As agencies across the United States continue to adopt the Balanced Mix Design (BMD) approach to evaluate the rutting and cracking resistance of asphalt mixtures, it is vital to thoroughly understand the variables that impact mixture performance test results (Elwardany *et al.* 2024, Mousavi Rad *et al.* 2025, Taylor 2025, Veginati *et al.* 2025). While significant research has focused on testing equipment variability, test method ruggedness, and specimen storage durations, ambiguity in BMD specimen fabrication standards continues to cause frustrating inter-laboratory differences (Hajj *et al.* 2022, Tiwari 2026). Because it is theorized that these procedural differences have a significant and practical impact on mechanical testing results, it is crucial to establish standardized practices so agencies and contractors can refine their procedures and reduce overall variability. To address these inconsistencies, a comprehensive evaluation of four critical areas of specimen fabrication variability, maximum specific gravity (G_{mm}) testing, specimen conditioning, oven heating protocols, and Hamburg Wheel Tracking (HWT) specimen preparation, is required.

To address the need for a practical and consistent evaluation of asphalt cracking, Im and Zhou (Im and Zhou 2017) introduced the indirect tensile asphalt cracking test (IDEAL-CT). Extensive laboratory calibration involving 200 tests demonstrated the method's sensitivity to critical variables, including binder content, volumetric properties (air voids), aging, and the inclusion of recycled materials such as reclaimed asphalt pavement (RAP). Typically conducted at 25℃ with a displacement rate of 50 mm/min per ASTM D8225, the IDEAL-CT index has shown strong empirical alignment with field performance data from MnRoad and the FHWA's Accelerated Load Facility (Zhou *et al.* 2017, ASTM International 2019).

To evaluate rutting distress in a laboratory setting, Wheel Load Tracking (WLT) tests are widely utilized, as they simulate the repetitive action of vehicular tires on HMA specimens. Specifically, the Hamburg Wheel Tracking Test (HWTT) follows the AASHTO T 324 protocol, where a 71.7 kg steel wheel passes over a submerged specimen at 50°C. This setup allows for continuous monitoring of vertical deformation relative to the number of load cycles (AASHTO 2023). While the HWTT assesses both moisture sensitivity and rutting through cyclic loading under water, the IDEAL-RT is a monotonic test focused on shear resistance. Both methodologies have become integral to the pavement industry, seeing increased adoption in state department of transportation (DOT) specifications and BMD protocols (Boz *et al.* 2022). Originally developed by the Texas A&M Transportation Institute, the IDEAL-RT was designed as a displacement-controlled test with a rate of 50 mm/min on samples conditioned at 50℃, according to ASTM D8360, for rapid shear strength evaluation (Zhou *et al.* 2019, 2020, ASTM International 2022). The IDEAL-RT's main advantage is its efficiency, delivering results within minutes versus the hours required for the HWTT and other repeated load tests. This makes it an ideal candidate for routine quality assurance (QA), with several studies showing a strong correlation with traditional rutting performance metrics (Rajan *et al.* 2023, Vamsikrishna and Singh 2023, Leavitt *et al.* 2024).

1.1 Maximum Specific Gravity (G_{mm}) Testing Variability (Task 1)

BMD protocols often require long-term oven aging (LTOA) conditioning to evaluate how aging influences the cracking performance of asphalt mixtures. For efficiency's sake, laboratories commonly use the short-term oven aging (STOA) G_{mm}—defined throughout this study for the plant-produced loose mixtures as reheated to compaction temperatures—to determine target air voids for specimens conditioned with LTOA (Smith 2017). This simplification raises an important question about whether, and to what extent, using STOA G_{mm} affects the actual air void content, which may, in turn, impact cracking test results. Literature shows that G_{mm} typically increases with long-term aging due to binder absorption into aggregates; as binder fills pore space, it displaces air and leads to a higher G_{mm} value (Lemke *et al.* 2018). Extended aging times allow more asphalt binder to absorb into the aggregates, altering the "true" G_{mm} and air void calculations. This emphasizes how critical it is to understand how well laboratory STOA protocols and further LTOA protocols simulate actual maximum specific gravity and binder absorption during plant production (Newcomb *et al.* 2015). It is hypothesized that mixtures with greater absorption potential will exhibit larger G_{mm} changes during extended oven aging. Further research is needed to assess whether changes in G_{mm} from aging significantly affect air void determination and, critically, the resulting BMD test outcomes. Additionally, it is important to determine how these effects vary depending on mixture composition.

1.2 Impact of Specimen Conditioning Methods (Task 2)

Another area lacking clear procedural standardization is the specific environmental conditioning method applied to specimens before performance testing. This research evaluates the effect of conditioning methods on the IDEAL-CT and IDEAL-RT tests as the ASTM standards allow both wet and dry conditioning. It is worth noting that while the standard conditioning time for the IDEAL-CT is identical for both wet and dry methods, the IDEAL-RT utilized significantly different durations depending on the environment: 150 minutes for dry conditioning in an environmental chamber versus only 45 minutes for wet conditioning in a water bath. Environmental chamber conditioning provides greater temperature control but requires larger, more expensive equipment and additional laboratory space. Water bath conditioning is inexpensive and convenient; however, most water baths used in asphalt testing lack active cooling, meaning the laboratory ambient temperature must be lower than the bath setpoint to achieve the target specimen temperature. In some cases, ice is added to lower the bath temperature, but this approach reduces the ability to maintain fine-tuned control. Despite these limitations, water baths are widely used in practice due to their low cost. Conditioning specimens in sealed plastic bags within a water bath (i.e., “dry” bath conditioning) is another option, but it presents practical challenges. For example, in a Virginia Transportation Research Council (VTRC) interlaboratory study, laboratories reported difficulty keeping specimens fully dry during bagged water bath conditioning (Habbouche *et al.* 2021).

Interlaboratory and comparative studies have produced mixed findings. The VTRC conducted a study involving five independent laboratories and two unique mixtures, finding that IDEAL-CT indices were not significantly affected by dry versus wet conditioning (Habbouche *et al.* 2022). In contrast, another research effort using two mixtures in a single laboratory reported slightly higher average CT index values for water bath conditioning compared to dry chamber conditioning, with the difference attributed to water ingress causing plastic deformation and creep rather than representing true failure energy (Kumar *et al.* 2024).

These divergent results highlight the need to further evaluate conditioning procedures to ensure reliable and comparable cracking performance evaluations. To address these uncertainties and better inform CAPRI members and the broader asphalt community, additional research incorporating a wider range of mixture types and component materials has been undertaken.

1.3 Impact of Oven Heating Protocols (Task 3)

The thermal history of a mixture, dictated by the oven heating protocol utilized in the laboratory, can significantly alter its aging profile and mechanical properties. Standardizing the reheating and aging procedure requires assessing factors such as oven type, sample size, and variations in reheating durations and temperatures (Lemke *et al.* 2018). A recent ruggedness study highlighted that mixture sample heating duration consistently demonstrated significant effects on the variability of both the CT index and high-temperature indirect tensile strength, emphasizing the need for strict control over heating practices (Boz *et al.* 2025). Furthermore, variations in reheating, conditioning temperatures, and hold times significantly impact the stiffness and cracking metrics of compacted specimens. Research indicates that LTOA conditioning temperature can have a greater effect on mixture aging than the LTOA duration itself, demonstrating how sensitive asphalt mixtures are to variations in thermal history (Newcomb *et al.* 2015). Evaluating different heating strategies, such as holding mixtures in preheated ovens for extended periods versus ramping temperatures up from cold ovens, is necessary to determine how deeply thermal history impacts BMD test outcomes.

1.4 Specimen Fabrication Impact on HWTT (Task 4)

Finally, the physical fabrication method of specimens prepared for the HWTT plays a crucial role in the accuracy of rutting evaluations. Specimens are typically either compacted directly to their final height (for example, 62 mm) or compacted taller and then trimmed to size. Compaction directly to 62 mm is generally the most commonly used approach; however, some aggregate structures will lock up before reaching 62 mm, making direct compaction to this height difficult. In these cases, compacting to a taller height such as 115 mm and then cutting is more feasible while still maintaining a final air void content of 7.0% ± 0.5% in the trimmed specimen. Recent evaluations of airfield asphalt mixtures noted that trimming specimens exposes the internal aggregate structure, which can lead to water infiltration during testing. Consequently, bottom-cut samples exhibited higher rut depths, increased variability, and early stripping failures compared to uncut samples (Hajj *et al.* 2025). This aligns with earlier findings that physical modifications and challenges involved in preparing field cores and cut specimens for testing molds, including aggregate degradation, debonding, and the use of plaster, introduce significant variability and can obscure actual rutting performance. These factors often cause field cores to correlate poorly with laboratory mixed, laboratory compacted specimens (Newcomb *et al.* 2015). Further validation across multiple mixtures and laboratories is needed to better understand the effects of trimming HWTT specimens to help mitigate fabrication-induced variability.

2 Objective and Scope

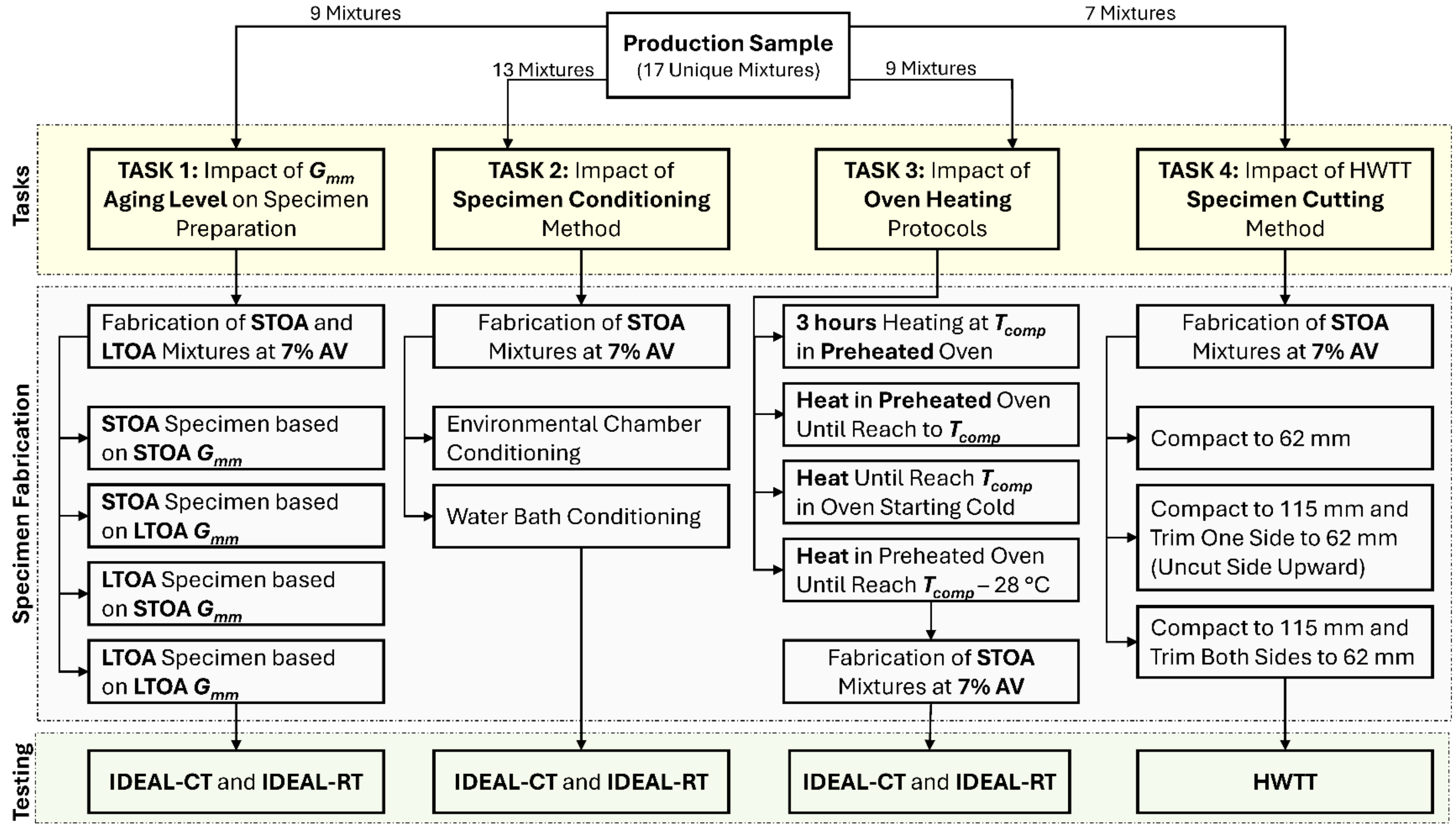
The primary objective of this project is to investigate the effects of variability in specimen fabrication procedures on BMD performance measures. The scope of the project examines four high-priority areas identified by the CAPRI technical evaluation subcommittee. The study will explore the impacts of G_{mm} variability, specimen conditioning, oven temperature, and the mode of preparation for HWTT specimens (cut vs. compacted specimens) to develop a comprehensive understanding of how these factors influence BMD results. To accomplish this, standardized test plans have been developed to investigate each of these four topics, allowing them to be accurately replicated for multiple mixtures and across various laboratories throughout the United States.

3 Testing Plan

3.1 Overview

Five laboratories were selected to execute the testing plan, with each lab contributing one or more unique mixtures and performing assigned tasks to evaluate specific fabrication variables; the experimental program was organized into four primary tasks, and an overall flow chart of the testing is provided in Figure 1. In total, 17 distinct mixtures were tested, representing a broad range of material properties (binder content, RAP content, G_{mm}, and aggregate absorption). Reheated loose plant mix was used for all portions of this study. For each mixture, the plant mix was sampled in a single batch using the participating agency's standard procedures.

Tasks 1–3 used the IDEAL-CT and IDEAL-RT to evaluate whether variations in specimen preparation and conditioning produced measurable differences in mixture performance; these tests were selected because of their relative simplicity and broad interest in BMD efforts nationwide. Task 4 evaluated the effect of cutting on HWTT specimens, a procedure with similarly broad use in BMD programs nationwide. For each mixture, G_{mm} was measured at the relevant aging conditions and verification pills compacted at design number of gyrations (N_{design}) were produced to confirm the mixtures matched the job mix formula (JMF) within acceptable tolerances.



T_{Comp} = Compaction temperature, AV = Air Voids

Figure 1. Test plan overview (source: FHWA).

3.2 Specimen Preparation

STOA was defined as the time required to reheat the loose plant mix, and laboratories were instructed to apply a consistent reheating procedure for all specimens. LTOA for all relevant tasks was 20 hours ± 30 minutes of loose mix oven aging at 110℃ in a pan (Zhou *et al.* 2025). Loose mix was placed in the pan at approximately 40 mm thickness, and no stirring was performed throughout the duration of the LTOA procedure.

For all specimens across all tasks (Including the specific oven aging protocols evaluated in Task 3), laboratories were instructed to stagger the timing of when loose mix was placed in the oven for compaction so that the difference in oven heating time for any batch did not exceed 60 minutes. The laboratories were further instructed that specimens were to be compacted to a height of 62 mm as quickly as practicable, and the elapsed time between compacting the first and last specimen in a given batch was not to exceed one hour. For example, if a laboratory placed enough mix in the oven to produce seven specimens, it must have been able to compact all seven specimens within one hour so that the time difference between the first and last specimen was ≤ 60 minutes. All specimens for a given task were required to be compacted on the same gyratory compactor. Parallel compacting was not permitted.

Specimens for IDEAL-CT were compacted to a height of 62 mm ± 1 mm with a target air-void content of 7.0% ± 0.5%; seven replicate specimens were tested per condition, conditioned at 25 ℃ (77 °F) using either wet or dry methods per ASTM D8225. Specimens for IDEAL-RT were also compacted to 62 mm ± 1 mm with a target air-void content of 7.0% ± 0.5%, with five replicate specimens per condition, conditioned at 50 ℃ (122 °F) using either wet or dry methods per ASTM D8363. HWTT specimens were either compacted directly to final dimensions or trimmed to size after compaction, with a final air-void content of 7.0% ± 0.5% and tested according to AASHTO T 324-23; four replicate HWTT tests (8 specimens total) were performed for each HWTT condition. Final air void values were measured and recorded for all specimens prior to testing.

3.3 Materials

The CAPRI Project Oversight Panel selected a total of 17 distinct mixtures from five participating laboratories to ensure a comprehensive evaluation. The geographic regions from which the selected mixtures originated are shown in Figure 2. The geographic diversity of interested laboratories was concentrated mostly in the Southern/Southeastern United States.

To capture a wide range of material properties, the selected mixtures incorporate diverse aggregate mineralogies, including granite, limestone, dolomite, traprock, chat, and gravel. All mixtures used either a 9.5 mm or 12.5 mm nominal maximum aggregate size (NMAS). Aggregate water absorption was a key factor in the selection process, with chosen mixtures representing a broad spectrum, from very low combined absorption (less than 0.5%) to highly absorptive aggregates exceeding 2.0%.

The experimental plan also introduces variability in binder grades and recycled material contents. Virgin asphalt binder performance grades (PG) span PG 52-28, PG 58-28, PG 64-22, PG 67-22, and PG 76-22. RAP content varies from 10% to as high as 40% by weight of the mixture. Several selected mixtures also contain chemical additives such as liquid anti-strip agents, warm-mix asphalt additives, and hydrated lime. Detailed mix design variables for all 17 mixtures are presented in Table 1.

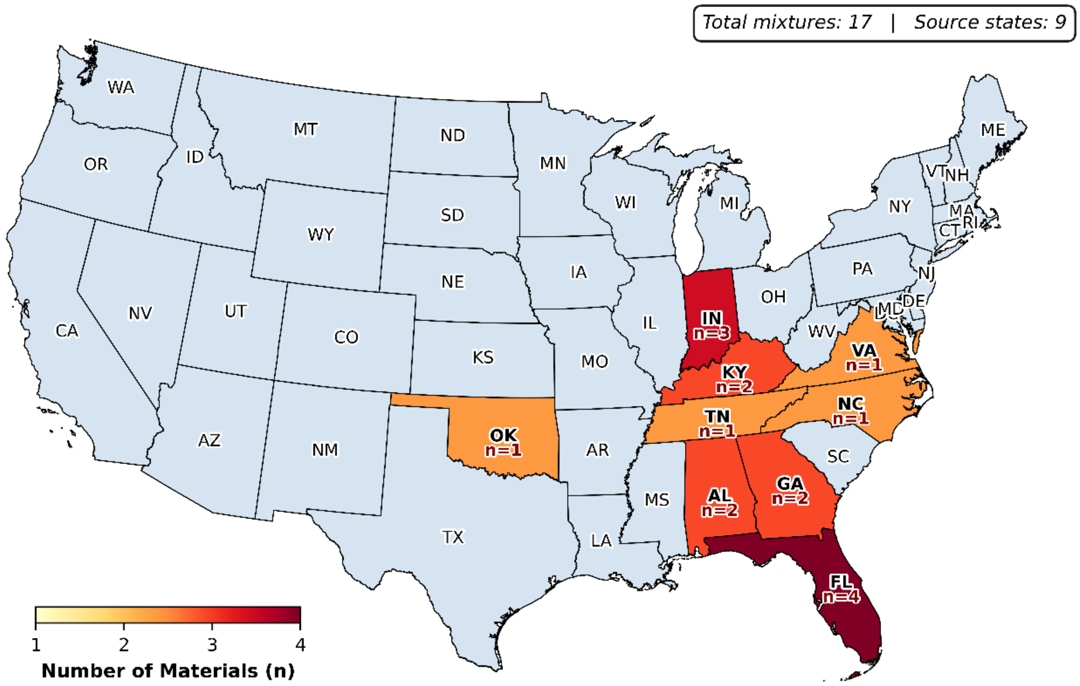


Figure 2. Origin map for mixtures used in this study (source: FHWA).

Table 1. Experimental mixture properties

Mixture ID	L1-M4	L1-M5	L2-M1	L2-M2	L2-M3	L2-M4	L3-M1	L3-M2	L4-M1	L4-M2	L4-M3	L4-M4	L4-M5	L4-M6	L4-M7	L4-M8	L5-M1
Design State	IN	IN	FL	FL	FL	FL	KY	KY	GA	AL	TN	AL	OK	GA	NC	VA	IN
NMAS (mm)	9.5	9.5	9.5	9.5	9.5	12.5	9.5	9.5	9.5	9.5	12.5	9.5	12.5	12.5	9.5	12.5	9.5
Binder PG	64-22	64-22	52-28	58-22	52-28	76-22	76-22	64-22	64-22	58-28	67-22	58-22	58-28	76-22	76-22	76-22	58-28
AC (%)	6.6	6.5	6.0	7.2	5.3	4.9	6.1	6.0	6.2	6.1	6.0	5.6	5.5	5.7	5.5	5.3	6.9
Virgin AC (%)	5.90	5.30	3.76	5.22	3.51	3.37	5.50	5.50	4.57	3.97	5.10*	3.64*	4.12*	3.99*	4.40*	4.50*	5.80
RAP ** **(%)**	10	20	40	30	35	30	10	10	30	35	15	35	25	30	20	15	20
G_{mm}	2.432	2.464	2.471	2.362	2.504	2.524	2.478	2.439	2.438	2.438	2.317	2.431	2.425	2.445	2.464	2.732	2.445
VMA (%)	17.0	17.0	16.0	15.7	15.6	15.0	15.4	16.9	16.7	16.1	15.7	14.9	15.7	15.5	15.8	17.3	16.8
VFA (%)	70.6	70.6	75.0	75.0	74.0	73.0	78.0	78.0	65.2	76.0	84.0	-	75.0	85.0	75.0	75.0	70.2
Dust/Pb	1.1	1.2	1.2	0.9	1.2	1.2	1.1	1.3	1.3	1.1	1.4	1.1	1.0	1.1	0.8	1.0	-
Water Abs %* **	1.77	2.07	1.59	2.36	0.77	0.68	1.78	1.25	0.69	0.72	3.07	0.66	1.65	0.57	0.88	0.45	1.77
Additive	-	-	AS	AS	AS	AS	-	-	AS	WMA	-	WMA	WMA	AS	WMA	WMA	-

*Approximate, not directly reported.
**RAP content is calculated by weight of the total mixture.
***Excluding RAP, since RAP absorption was not measured.

3.4 Task Details

Task 1: This experiment compared the effect of using STOA versus LTOA G_{mm} values when preparing BMD specimens. Because cracking tests often require LTOA conditioning, it is common practice to use the STOA G_{mm} to set target air voids for both STOA- and LTOA-conditioned specimens to reduce complexity. Task 1 evaluated whether that simplification significantly changes the actual air voids and, consequently, the cracking test results. The performance specimen fabrication process was split into two parallel tracks, with all specimens targeted to a height of 62 mm and an air void content of 7% ± 0.5%:

- Using the STOA G_{mm} for air void determination: laboratories compacted 12 STOA specimens (7 for IDEAL-CT and 5 for IDEAL-RT) alongside 7 LTOA specimens (for IDEAL-CT).
- Using the LTOA G_{mm} for air void determination: laboratories compacted an identical set of 12 STOA specimens (7 for IDEAL-CT and 5 for IDEAL-RT) alongside 7 LTOA specimens (for IDEAL-CT).

All specimens were conditioned and tested within three days of compaction. Following conditioning, all specimens were tested immediately according to ASTM D8225 (IDEAL-CT) and ASTM D8363 (IDEAL-RT) protocols.

Task 2: This experiment evaluated the effect of specimen conditioning methods (dry versus wet) on IDEAL-CT and IDEAL-RT performance test outcomes. Thirteen mixtures provided by all five participating laboratories were tested, producing 178 IDEAL-CT records and 126 IDEAL-RT records. Specimens were fabricated to a target height of 62 mm and an air void content of 7% ± 0.5%, using a Superpave gyratory compactor. For each mixture the performance specimen conditioning process was divided into two parallel tracks:

Dry Conditioning: specimens were placed in an environmental chamber and held at the specified test temperature for the required time.

- IDEAL-CT specimens were conditioned at 25°C (77°F) for 2 hours ± 10 minutes.
- IDEAL-RT specimens were conditioned at 50°C (122°F) for 150 ± 10 minutes.

Wet Conditioning: specimens were submerged directly in a clean water bath (unbagged), maintained at the specified test temperature.

- IDEAL-CT specimens were conditioned at 25°C (77°F) for 2 hours ± 10 minutes.
- IDEAL-RT specimens were conditioned at 50°C (122°F) for 45 ± 5 minutes.

All specimens were conditioned and tested within three days of compaction. Following conditioning, specimens were tested immediately according to ASTM D8225 (IDEAL-CT) and ASTM D8363 (IDEAL-RT) protocols.

Task 3: This experiment investigated the effect of four distinct loose mix oven heating protocols on IDEAL-CT and IDEAL-RT outcomes to determine how thermal history during specimen fabrication influences measured mixture performance. Eleven mixtures from four participating laboratories were evaluated, producing 284 IDEAL-CT records and 199 IDEAL-RT records. Specimens were fabricated to a target height of 62 mm and an air void content of 7.0% ± 0.5% and tested within three days of compaction. Prior to compaction, bulk loose mix was heated only until it became workable, then subdivided into small batches approximately equal to the target compaction weight. Batches were cooled back to laboratory ambient temperature before proceeding with the designated heating protocol for compaction. Four methods were investigated to represent the range of fabrication procedures used in practice:

- Method A — Preheated hold: the oven was preheated to the compaction temperature, and the loose mix was placed in the preheated oven for 3 hours prior to compaction.
- Method B — Direct heat to temperature: the loose mix was heated until it reached the compaction temperature ± 3℃ (± 5°F), then removed for compaction. Ovens were permitted to be overheated by no more than 5.6℃ (10°F).
- Method C — Cold oven ramp: the loose mix was placed in a cold oven, the oven was turned on and set to temperature, and the mix was heated in situ until it reached the compaction temperature ± 3℃ (± 5°F). Oven set temperature was permitted to be set no more than 5.6℃ (10°F) above the compaction temperature.
- Method D — Under-heated: the loose mix was heated in an oven set to the compaction temperature until it reached a temperature 25 ± 3℃ (45 ± 5°F) below the compaction temperature then immediately compacted.

After the assigned heating protocol, specimens were tested with IDEAL-CT at 25℃ (77°F) and IDEAL-RT at 50℃ (122°F) following the referenced ASTM procedures (IDEAL-CT per ASTM D8225; IDEAL-RT per ASTM D8363). The objective of Task 3 was to quantify whether and how different reheating histories affect cracking related test metrics and to identify any procedural controls needed for consistent BMD specimen preparation.

Task 4: This experiment investigated how physical fabrication methods affect HWTT results to determine whether trimming introduces measurable variability. Seven mixtures from all five participating laboratories were evaluated, producing 84 HWTT records. Specimens were prepared to a target height of 62 mm and a final air void content of 7.0% ± 0.5% and tested within 10 days of compaction. For each mixture, specimens were tested using three fabrication methods:

- Method A — Direct compaction to final height: specimens were compacted directly to a final height of 62 mm.
- Method B — Single side trimmed: specimens were compacted to 115 mm and then trimmed on one side to achieve the 62 mm final height. Testing was performed with the uncut side of the specimen facing up.
- Method C — Double side trimmed: specimens were compacted to 115 mm and trimmed on both sides to achieve the 62 mm final height.

Four replicate HWTT specimens were produced per fabrication method for each mixture (four replicate tests per method per mixture, yielding 24 specimens per mixture and 168 total specimens). Final specimen heights and air void contents were recorded prior to testing. All HWTT testing was conducted at 50℃ (122°F) and in accordance with AASHTO T 324-23.

4 Results and Analysis

This section presents the results and discussions derived from experimental and statistical evaluations to assess the effect of different specimen fabrication methods on BMD performance test results. This section is structured into four subsections; each focused on one task.

4.1 Task 1: Effect of G_{mm} Aging Level

Task 1 was designed to evaluate the specific influence of G_{mm} aging protocols on the performance metrics derived from the IDEAL-CT and IDEAL-RT tests. For this purpose, nine distinct plant-produced asphalt mixtures were collected and subsequently evaluated by five independent asphalt laboratories across the United States. Figure 3 presents the G_{mm} values for these mixtures under both STOA and LTOA aging conditions. As illustrated in the figure, the LTOA protocol generally resulted in an increased G_{mm} relative to the STOA baseline, which is primarily attributed to two concurrent mechanisms during extended heating: the volatilization of light oily fractions and the progressive oxidation of the asphalt binder, which inherently hardens the binder and increases its density. Furthermore, prolonged exposure to elevated temperatures facilitates continued asphalt absorption into the aggregate pore structure, marginally reducing the effective asphalt content and driving up the G_{mm} of the mixture (Cooley Jr and Williams 2013, Pasandín and Pérez 2014). Figure 3 showed that the aging effect on measured G_{mm} values was largely marginal across the dataset. With the explicit exception of mixture L2-M2, the transition from STOA to LTOA produced very slight G_{mm} variations, yielding maximum increases of only 0.52%. In stark contrast to the broader dataset, mixture L2-M2 exhibited a distinctly higher sensitivity to the aging level and recorded a pronounced G_{mm} increase of 4.68%. It is also noted that the difference between the STOA and LTOA G_{mm} values were lower than the acceptable range of Gmm measurements based on multi-laboratory precision estimation in AASHTO T 209-25 (AASHTO 2025).

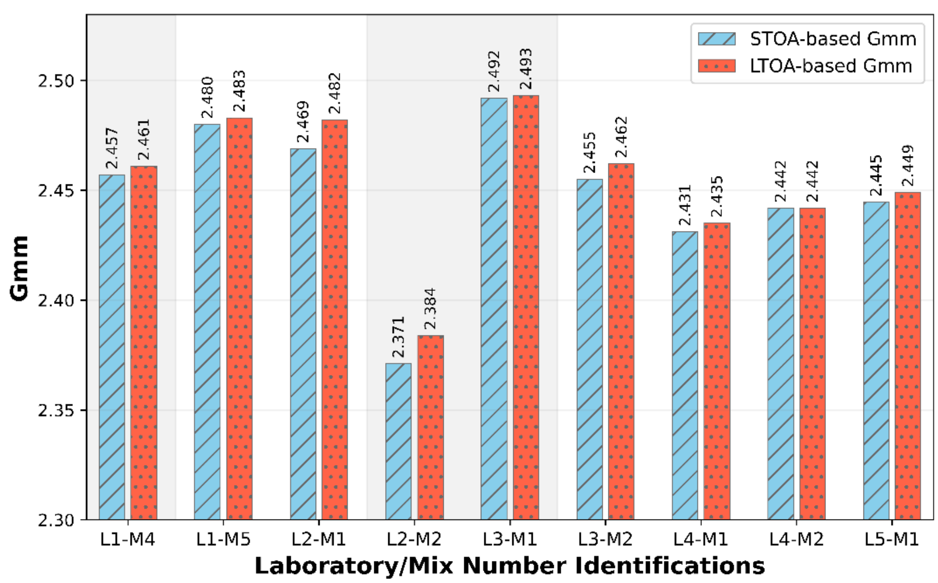


Figure 3. Reported G_{mm} values for each mixture at STOA and LTOA aging levels (source: FHWA).

To contextualize the practical implications of these density shifts, the effect of the G_{mm} aging level on the calculated volumetric properties of the test specimens was also evaluated, as detailed in Appendix 3 (see page 57). Because the G_{mm} of an LTOA mixture is generally higher than that of a STOA mixture, applying the STOA G_{mm} to LTOA specimens underestimates the true air void content. Across all nine mixtures evaluated, substituting the STOA G_{mm} for the true LTOA G_{mm} resulted in calculated specimen air voids being artificially lower by an average of 0.2%, with a maximum observed discrepancy of 0.5% in highly absorptive mixtures. This discrepancy carries significant practical importance for laboratory fabrication, as standard test tolerances are typically tightly constrained to a 1.0% window (e.g., 6.5% to 7.5%). Consequently, when practitioners utilize the more readily available STOA G_{mm} to compact LTOA specimens, it would be conservative to target the upper half of the acceptable air void range (e.g., 7.0% to 7.5%). This proactive adjustment ensures that the true specimen air voids do not inadvertently drop below the 6.5% minimum threshold due to the reference G_{mm} underestimation.

Following the volumetric assessment, the core evaluation of Task 1 focused on the measured performance metrics from IDEAL-CT and IDEAL-RT tests. For this purpose, the RT- and CT-indices, as the main output parameters of these two BMD performance tests, were plotted for all evaluated mixtures of this task in Figure 4 and Figure 5, respectively.

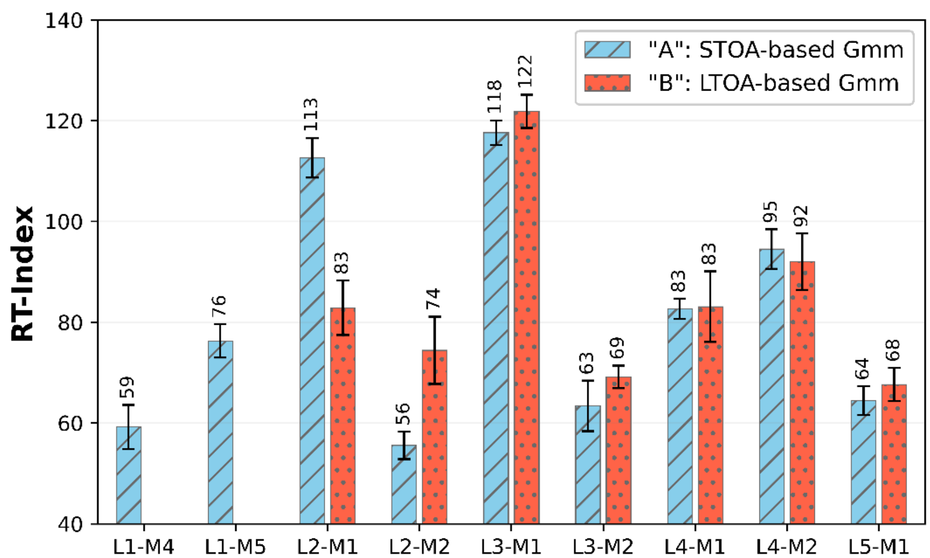


Figure 4. Summary of calculated RT-indices for all evaluated mixtures in task 1 (source: FHWA).

As shown in Figure 4 for the IDEAL-RT results, the selected plant-produced mixtures exhibit a wide range of measured rutting performance, with mixtures exhibiting average RT-indices spanning from 56 to 122. Similarly, the distribution of the IDEAL-CT results in Figure 5 showed an average CT-indices as low as 15, scaling continuously to more crack resistant mixtures with indices reaching up to 158. This range in both performance indices indicates that the database encompasses a broad, representative cross-section of mixture behaviors which helps ensure that the subsequent statistical evaluation of the G_{mm} aging effect is relevant across a wide variety of mixtures.

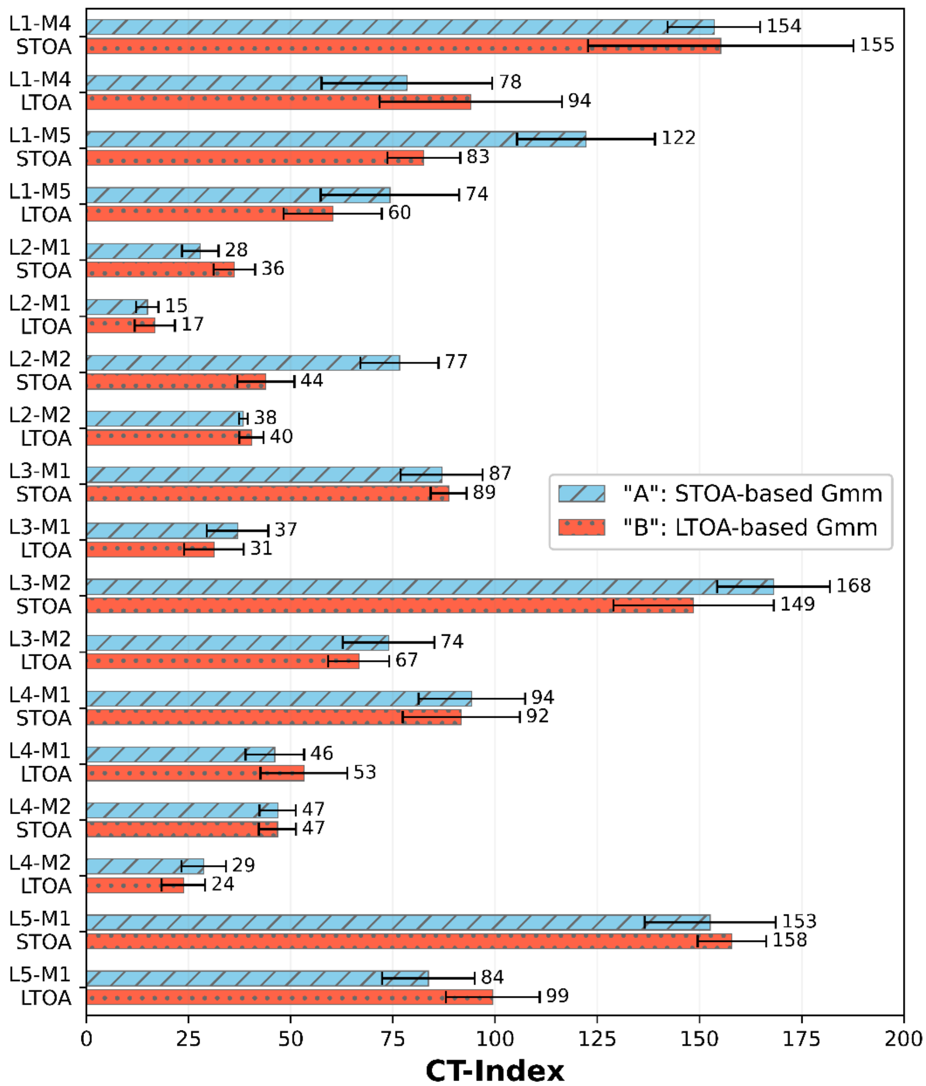


Figure 5. Summary of calculated CT-indices for all evaluated mixtures in task 1 (source: FHWA).

Given the availability of the raw testing data, a suite of performance metrics was calculated for each IDEAL-CT and IDEAL-RT test replicate. For the IDEAL-RT, these outputs included the RT-index and tensile strength. For the IDEAL-CT, the extracted parameters included the CT-index, flexibility parameter (l_{75}/m_{75}), failure energy (G_f), Cracking Propagation Resistance (CPR), and tensile strength. Detailed methodologies for these parameters are documented in Appendix 2 (see page 52).

To initially visualize the influence of the G_{mm} aging level on these performance metrics, equality plots were generated comparing the indices of specimens compacted to targets based on STOA versus LTOA G_{mm}. Figure 6 shows this comparative approach using the CT-index parameter as a representative example. As shown in this figure, the plot consists of 18 distinct data points—representing the nine mixtures evaluated at both STOA and LTOA conditions— with error bars indicating the intra-set standard deviation.

To quantify the maximum observed discrepancy introduced by the varying G_{mm} targets, a bounding deviation envelope (shaded region) was applied to the data. This envelope is established by the single data point exhibiting the greatest distance from the line of equality, covering the entire dataset. The upper and lower limits of this bounding envelope are defined by Equation 4.1.

$$y_{upper} = (1 + Max\ Deviation) \times x$$
$$y_{lower} = \left(\frac{1}{1 + Max\ Deviation}\right) \times x \quad (4.1)$$

Where *y* and *x* represent the parameter values on the vertical and horizontal axes, respectively, and the *Max Deviation* is the fractional discrepancy dictating the spread of the envelope. Functionally, this envelope serves as a non-statistical, deterministic boundary that defines the maximum percentage difference required to assume parity between the two axes. For the CT-index dataset, the bounding envelope was calculated at 74.4%. From a practical perspective, this indicates that the absolute worst-case discrepancy caused by the G_{mm} aging protocol was 74.4%; therefore, to universally disregard the effect of the G_{mm} aging level on IDEAL-CT results, one would expect a potential parameter variation of up to 75%.

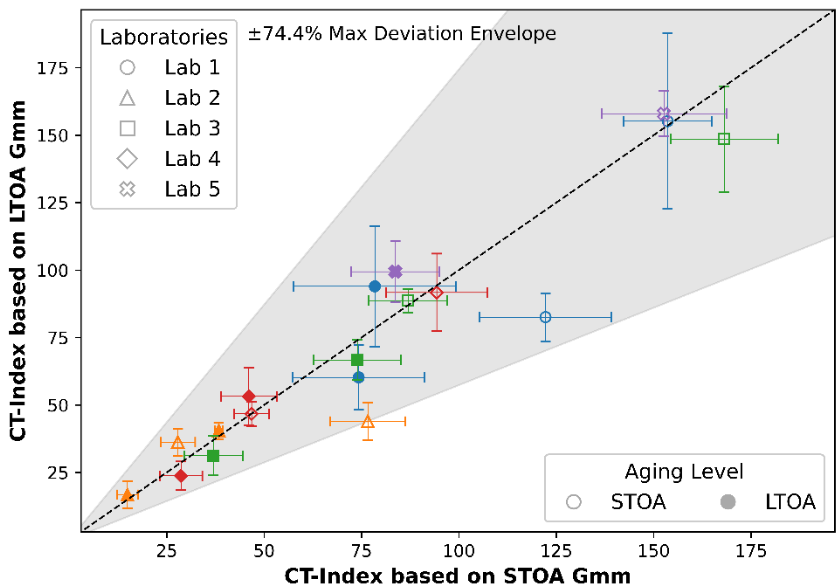


Figure 6. Equality plot for CT-indices of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

In the next step, Student's t-tests were conducted on each corresponding sample set to evaluate the differences between specimens compacted to STOA versus LTOA G_{mm} targets. The calculated t-statistics and corresponding p-values from these analyses were plotted on p-value functionality plots, where an example of these plots for CT-index parameter is shown in Figure 7.

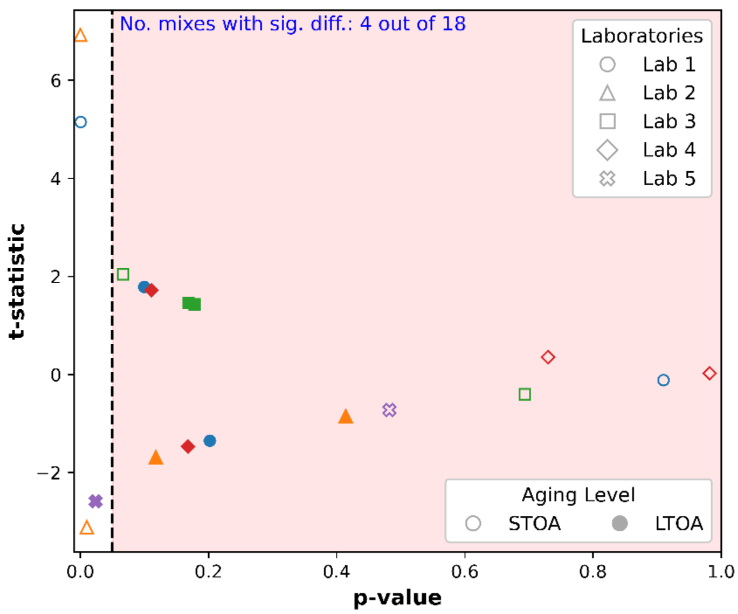


Figure 7. p-value functionality plot for CT-index parameter in task 1 (source: FHWA).

As shown in Figure 7 for the CT-index parameter, the majority of the mixtures yielded p-values greater than the $\alpha = 0.05$ threshold, which indicates that the impact of the G_{mm} aging level was statistically insignificant for those specific mixtures. Only four of the eighteen evaluated mixtures exhibited a statistically significant shift in their calculated CT-indices due to the G_{mm} aging level. Furthermore, among these four significant cases, the calculated t-statistics revealed a divergent directional response: two mixtures yielded positive t-statistics, while the remaining two were negative. This split confirms that the G_{mm} aging protocol does not exert a uniform, unidirectional influence meaning it does not consistently increase or consistently decrease the cracking index across the board. This mixture-dependent behavior is clearly illustrated by comparing the responses of specific sample sets. For instance, when evaluating mixture L5-M1 at the LTOA condition, utilizing the LTOA G_{mm} compaction target significantly increased the mean CT-index from 84 to 99. Conversely, when evaluating mixture L1-M5 at the STOA condition, the application of the LTOA G_{mm} target caused a significant reduction in the CT-index, dropping the mean value from 122 to 83.

While the equality and p-value functionality plots for the CT-index parameter are presented above, the corresponding plots for the remaining performance parameters are compiled comprehensively in Appendix 3 (see sections A3.2 and A3.3). To synthesize these findings, Table 2 summarizes the calculated maximum deviation envelope limits alongside the aggregate Student’s t-test results for all evaluated parameters.

Table 2. Summary of the Student's t-test for mixtures in task 1.

Parameter	Max. Deviation (%)	Total Number of Sample Sets*	Number of Sample Sets with Significant Difference (p-value < 0.05)		
			All t-statistics**	**t-statistic > 0**	**t-statistic < 0**
CT-Index	74.4	18	4 (22.2%)	2	2
RT-Index	35.8	7	3 (42.9%)	1	2
Flexibility	58.7	18	6 (33.3%)	4	2
G_f	10.1	18	1 (5.6%)	1	0
CPR	18.3	18	1 (5.6%)	1	0
Tensile Strength	9.6	18	5 (27.8%)	1	4

*Total number of the corresponding sample sets which were compacted using similar mixture at the same aging level but targeting STOA and LTOA G_{mm} aging levels.

**Percentages were calculated and presented in parenthesis.

As shown in Table 2, the maximum deviation envelopes across the evaluated parameters ranged widely from 9.6% to 74.4%. Notably, parameters representing more fundamental physical measurements, such as tensile strength and G_f, exhibited significantly tighter deviations. In contrast, composite metrics like the CT-index and flexibility parameter showed considerably higher maximum deviations. However, the mixture-specific statistical analysis utilizing the Student's t-tests revealed that there is no correlation between the deterministic maximum deviation and the actual statistical significance of the G_{mm} aging effect. For instance, while tensile strength exhibited the lowest maximum deviation (9.6%)—indicating that the sample means clustered closely around the line of equality—the t-tests still identified a statistically significant G_{mm} aging effect in 27.8% of the evaluated mixtures. Conversely, the flexibility parameter yielded a comparable rate of statistical significance (33.3%) but was characterized by one of the highest maximum deviations at 58.7%. Furthermore, a critical observation from this summary is the divergent directionality of the significant results. For nearly all parameters where more than one sample set was significantly affected by the G_{mm} aging level, the performance shift was not unidirectional. The concurrent presence of both positive and negative t-statistics across the dataset indicates that if the G_{mm} aging protocol does induce a significant change, its directional impact is mixture-dependent.

To assess the simultaneous impact of the experimental variables across the entire database, a two-way ANOVA was conducted. Table 3 presents the detailed two-way ANOVA output for the CT-index parameter, utilizing the robust HC3 standard error methodology detailed in Appendix 1 to account for the dataset's variance characteristics. The two-way ANOVA results presented in this table revealed two main outcomes. First, the main effect of the specific mixture type yielded a highly significant p-value well below the 0.05 threshold ($p = 0.00$). This makes intuitive sense and indicates that the database includes distinctly different mixtures with different material combinations and properties that dictate their cracking performance. The average database CT-indices ranges from approximately 15 to 158. On the other hand, the main effect of the G_{mm} aging level was found to be statistically insignificant ($p = 0.31 \gg 0.05$). From a macro-level perspective, this indicates that applying the LTOA G_{mm} target does not induce a universal, systematic change in the cracking resistance across the broad spectrum of mixtures.

Table 3. Results of two-way ANOVA based on CT-Index parameter for Task 1.

Parameter	**SSE**	**Degree of Freedom**	**F-statistic**	**p-value**
G_{mm} Aging Level	888.32	1	1.029	0.31
Mixture	442,013.08	8	63.986	0.00
Combined Effect	15,677.39	8	2.269	0.02
Residual	195,151.23	226		

Table 3 also showed that the two-way ANOVA model yielded a statistically significant interaction effect ($p = 0.02$), indicating a non-uniform response. While the G_{mm} aging protocol fails to show a significant overall effect on the generalized dataset, this significant interaction statistically demonstrates that the impact of the G_{mm} aging level on the CT-index parameter is mixture-dependent. This finding aligns with the preceding Student's t-test results, which showed that the specific magnitude and direction of the performance shift caused by the G_{mm} aging level depends on the unique characteristics of the individual asphalt mixture.

The statistical findings from two-way ANOVA analysis across all evaluated performance metrics and mixtures were aggregated and summarized in Table 4. A review of these results highlights several distinct behavioral trends regarding how different parameters respond to the experimental variables.

Table 4. Summary of two-way ANOVA results for IDEAL-CT and IDEAL-RT tests in Task 1.

Parameter	**G_{mm} Aging Level Effect**		**Mixture Effect**		**Combined Effect**	
	F-statistic	**p-value**	**F-statistic**	**p-value**	**F-statistic**	**p-value**
CT-Index	1.029	✗ 0.31	63.986	✓ 0.00	2.269	✓ 0.02
RT-Index	71.425	✓ 0.00	304.762	✓ 0.00	29.877	✓ 0.00
Flexibility	1.227	✗ 0.27	58.706	✓ 0.00	2.204	✓ 0.03
Failure Energy	2.708	✗ 0.10	140.616	✓ 0.00	1.147	✗ 0.33
CPR	1.766	✗ 0.19	37.656	✓ 0.00	0.623	✗ 0.75
Tensile Strength	6.254	✓ 0.01	71.532	✓ 0.00	1.083	✗ 0.38

First, the main effect of the mixture type is found highly significant across all evaluated parameters. From an engineering perspective, this is an expected and validating outcome. It confirms the extent of broad dataset of mixture and the fact that these tests and their output parameters are sensitive to different mixtures with various material properties.

Second, the results showed that the RT-index proved to be the most sensitive parameter in the dataset, standing as the only metric where all three factors are statistically significant. This outcome indicates dual sensitivity: the G_{mm} aging level alters the measured RT-indices across the broader dataset (the significant main effect), but the specific magnitude and severity of that alteration are mixture dependent.

When evaluating the failure energy and the CPR parameters, the analysis reveals that only the mixture effect is significant. Neither the G_{mm} aging level nor the combined interaction term were statistically significant. This indicates that these parameters are unaffected by variations in the G_{mm} aging target, either on a macro-level or a mixture-specific basis. In case of the CPR parameter, such behavior likely stems from the relatively low signal-to-noise ratio inherent to the metric, as it visually evident in the equality plot presented in Appendix 2 (see Figure A3.6). On the other hand, the flexibility parameter perfectly mirrored the behavior of the CT-index discussed previously: the main effect of G_{mm} aging was insignificant, but the combined interaction effect was significant. This reiterates that for these composite cracking indices, the influence of the G_{mm} aging level is entirely mixture-dependent rather than a uniform global shift.

Finally, a distinctly different statistical phenomenon is observed with tensile strength. For this parameter, both the main effect of the mixture and the main effect of the G_{mm} aging level are significant, while the combined interaction effect is not. The lack of a significant interaction term is critical; it indicates that the G_{mm} aging level exerts a consistent, uniform shift on the tensile strength across all evaluated mixtures. Specifically, since LTOA conditioning inherently raises the G_{mm} of the mixtures, utilizing these higher LTOA G_{mm} values during the compaction increases the measured tensile strength of the specimens. In other words, altering the G_{mm} target impacts the tensile strength of the specimens, and this directional influence occurs independently of the specific mixture type.

In a separate analysis, the correlations between mixture water absorption and performance metrics derived from the IDEAL-CT and IDEAL-RT tests were examined. The results revealed no discernible trend for the IDEAL-CT-related performance metrics; however, a relatively weak but notable trend was identified for the RT-index, as shown in Figure 8. In this figure, the relative difference in RT-index was computed by comparing the average RT-indices of specimens compacted using the LTOA G_{mm} against those compacted using the STOA G_{mm} as the reference. As shown, despite a relatively low coefficient of determination (R^2 = 21.7%), the trend suggests that mixtures with higher water absorption may exhibit a greater sensitivity of the RT-index to the G_{mm} aging level used during compaction. If this trend holds true, using the STOA G_{mm} to compact LTOA specimens with highly absorptive aggregates may result in a slightly lower RT-Index than if the LTOA G_{mm} were used.

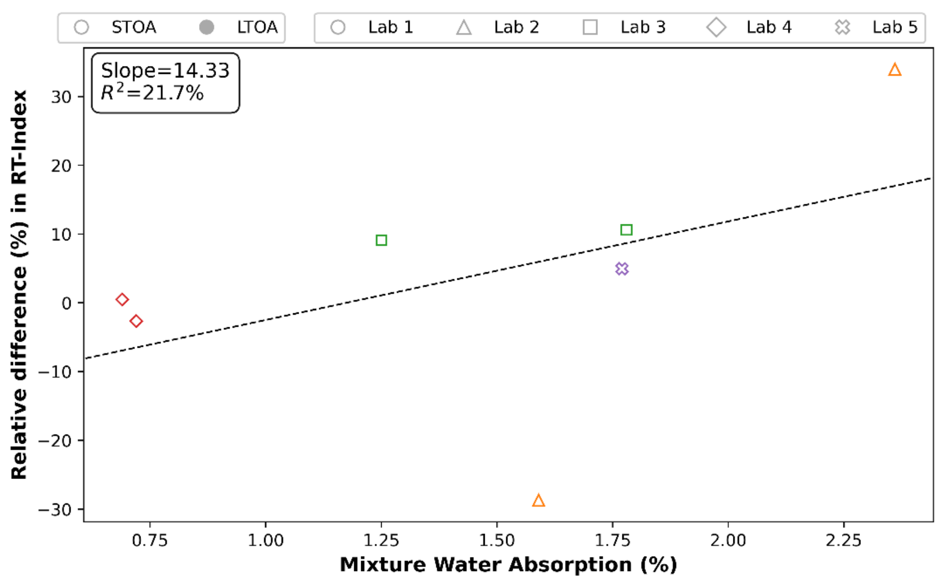


Figure 8. Correlation between mixture water absorption and relative difference in RT-index (source: FHWA).

To further investigate this phenomenon, the correlation between mixture water absorption and the relative difference in G_{mm} values was also evaluated, as shown in Figure 9. The results indicate that the relative difference in measured G_{mm} values across aging levels increases with increasing water absorption, particularly for mixtures with water absorption values exceeding 2%. This higher relative difference corresponds to a denser compacted specimen — that is, targeting a 7% air void content using the higher G_{mm} obtained at the LTOA aging level results in specimens with lower actual air voids compared to those compacted using the lower STOA G_{mm} value. Nevertheless, the results presented in Appendix 3 (see page 57) indicate that the G_{mm} aging level affects final air void content by a maximum of approximately 0.5%. This magnitude is insufficient to consistently influence IDEAL-CT parameters, given that the inherent variability of the IDEAL-CT test already exceeds this range. In contrast, the IDEAL-RT test, which exhibits comparatively lower testing variability, is sensitive enough to detect this difference, giving rise to the weak but discernible trend observed in Figure 8.

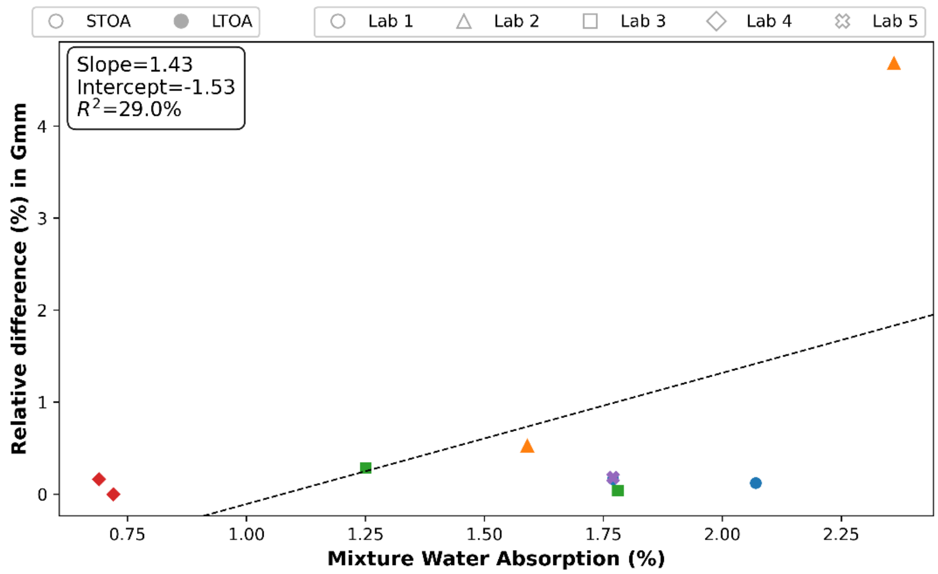


Figure 9. Correlation between mixture water absorption and relative difference in G_{mm} (source: FHWA).

4.2 Task 2: Effect of Conditioning Method

Task 2 was designed to evaluate the specific influence of the pre-test conditioning method on the performance metrics derived from the IDEAL-CT and IDEAL-RT tests. The experimental conditioning protocols evaluated in this task are primarily categorized into two distinct classes: (a) wet conditioning, which involves submerging the test specimens in a water bath for either 120 minutes at 25℃ for the IDEAL-CT or 45 minutes at 50℃ for the IDEAL-RT; and (b) dry conditioning, which requires placing the specimens within a temperature-controlled environmental chamber for 120 minutes at either 25℃ or 50℃ for IDEAL-CT and IDEAL-RT tests, respectively.

In this task, thirteen distinct plant-produced asphalt mixtures were collected and subsequently evaluated by five independent asphalt laboratories across the United States. As detailed in the supporting volumetric analysis provided in Appendix A4.1, all laboratory fabricated test specimens were successfully compacted to the target air voids of 7.0% ± 0.5%. Furthermore, a supplementary analysis (Appendix A4.2) showed that the standard sequential testing workflow utilized by these laboratories did not introduce a measurable testing bias. Specifically, no significant variations were observed in the measured CT- and RT-indices among replicates within a single sample set, that were tested within 16 minutes between the first and final tests.

To evaluate the specific impact of the conditioning method, this task utilized the IDEAL-RT and IDEAL-CT tests as the primary performance baselines. Accordingly, the RT- and CT-indices, as the primary output parameters of these tests, are plotted for all evaluated mixtures in Figure 10 and Figure 11, respectively.

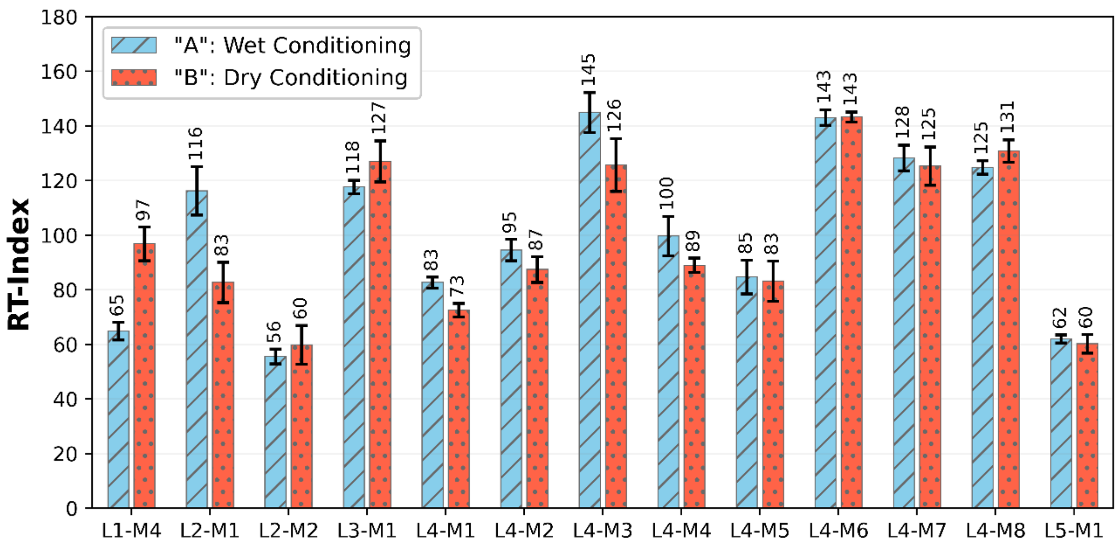


Figure 10. Summary of calculated RT-indices for all evaluated mixtures in task 2 (source: FHWA).

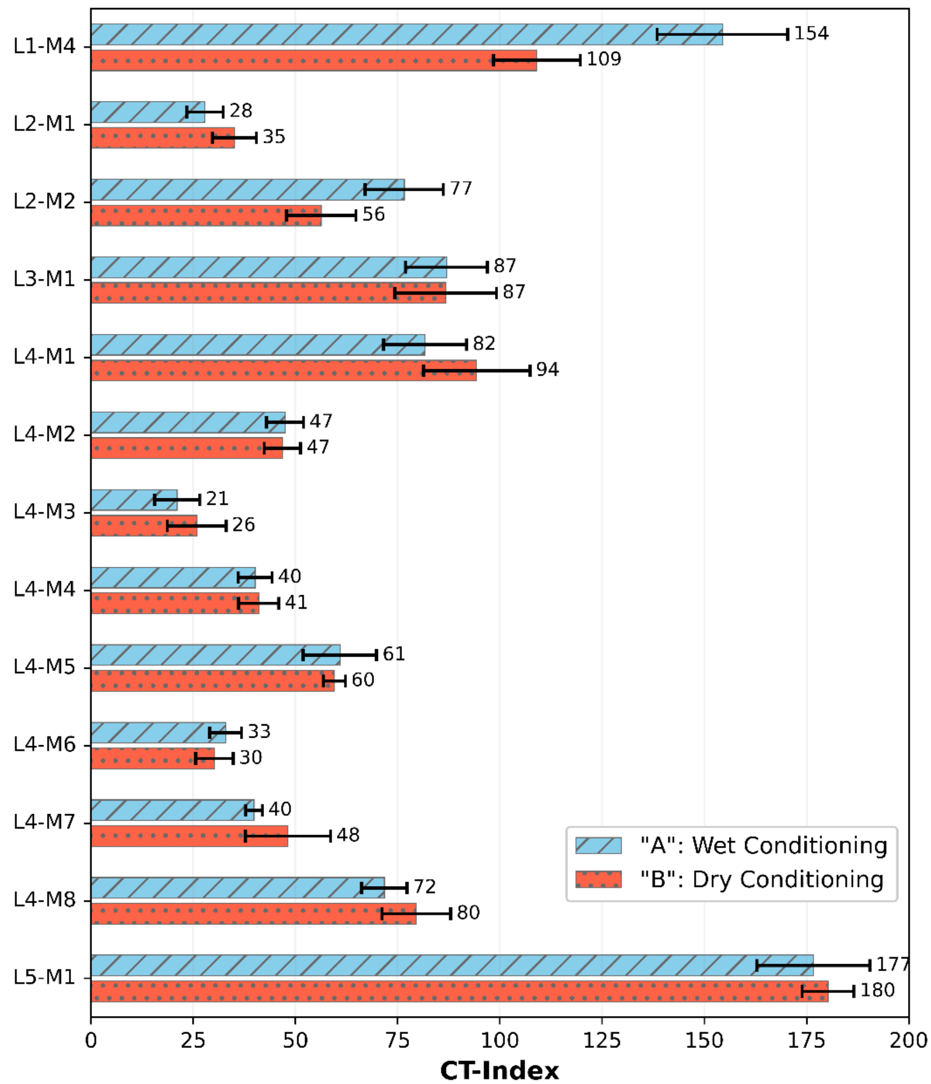


Figure 11. Summary of calculated CT-indices for all evaluated mixtures in task 2 (source: FHWA).

As illustrated in Figure 11, the CT-indices indicate that the selected plant-produced mixtures represent a robust spectrum of cracking performance. The dataset captures mixtures exhibiting average CT-indices as low as 21 and up to 180. Similarly, the distribution of the RT-indices in Figure 10 demonstrates a good spread in measured performance, with average RT-indices ranging from 56 to 146. This range across both performance metrics confirms that the assembled database includes a broad, representative cross-section of mixture behaviors. This ensures that the statistical evaluation of the conditioning methods includes a wide variety of typical mixture performance.

Similar to the methodology employed in Task 1, a suite of performance metrics was calculated from the raw testing data files. For the IDEAL-RT, these outputs included the RT-index and tensile strength. For the IDEAL-CT, the extracted parameters included the CT-index, Flexibility parameter, G_f, CPR, and tensile strength. Detailed methodologies outlining the calculation of these parameters are documented in Appendix 2.

The effect of the conditioning method was initially visualized utilizing equality plots, which compare the various performance metrics calculated for sample sets subjected to the wet versus dry conditioning protocols. As an example, Figure 12 shows this comparative approach using the CT-index parameter. As this figure shows, the plot consists of 13 distinct data points, each representing a specific mixture within the dataset, and error bars indicating the intra-set standard deviation. The shaded region on the equality plot delineates the maximum deviation envelope, the theoretical details of which were established previously in section 4.1 (see page 14).

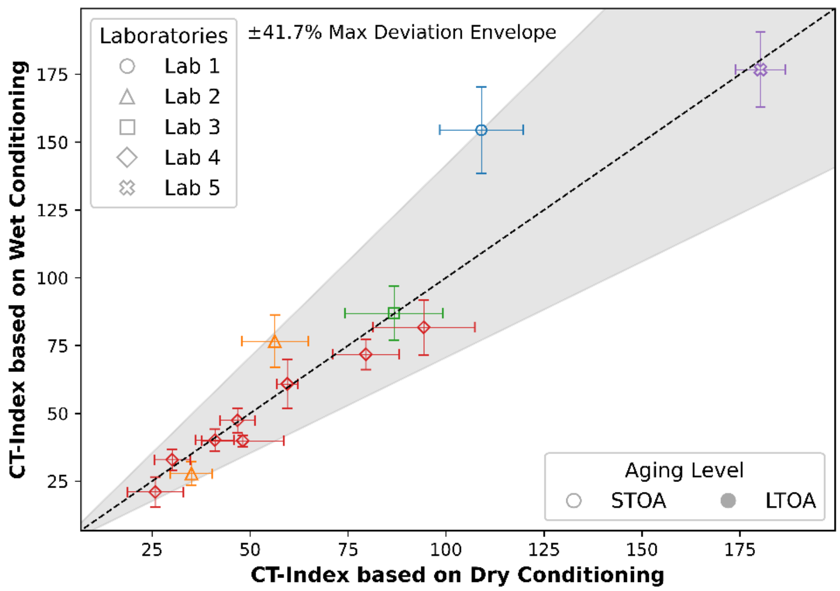


Figure 12. Equality plot for CT-indices of specimens compacted based on wet and dry conditioning (source: FHWA).

As previously discussed, this envelope serves as a non-statistical, deterministic boundary that defines the maximum percentage difference required to assume parity between the two axes. For the CT-index dataset, the bounding envelope was calculated at 41.7%. From a practical perspective, this translates to an absolute worst-case discrepancy of 41.7% caused strictly by the choice of conditioning method. Therefore, to universally disregard the effect of the conditioning method on IDEAL-CT results, one would have to accept a potential error of over 40%.

To perform a statistical evaluation of the conditioning method's effect on each individual mixture, Student's t-tests were conducted on each corresponding sample set. The calculated t-statistics and corresponding p-values from these analyses for CT-index parameter were plotted in a p-value functionality plot, as shown in Figure 13. As this figure shows, ten mixtures yielded p-values greater than the standard $\alpha = 0.05$ threshold, and only three mixtures exhibited a statistically significant shift in their calculated CT-index due to the conditioning method ($p < 0.05$). Furthermore, among these three significant cases, the calculated t-statistics revealed a divergent directional response: one mixture yielded a positive t-statistic, while the remaining two were negative. This split confirms that the conditioning method does not exert a uniform, unidirectional influence across the board.

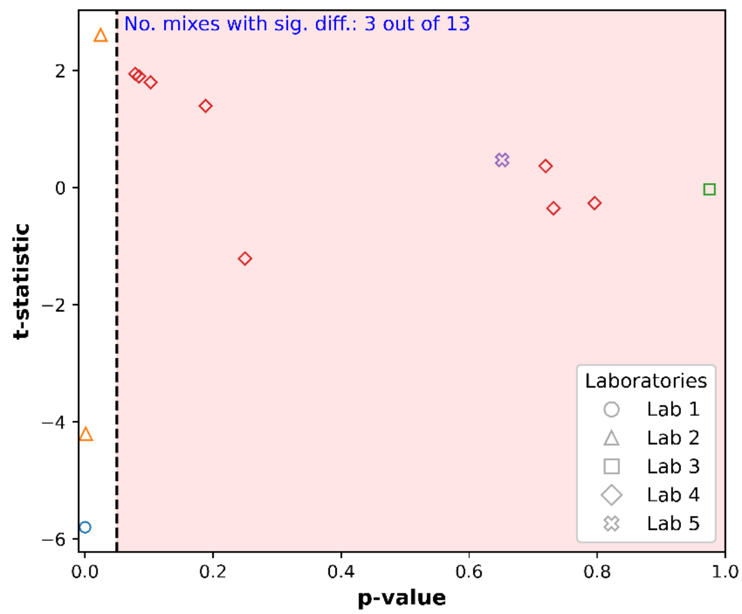


Figure 13. p-value functionality plot for CT-index parameter in task 2 (source: FHWA).

While the equality and p-value functionality plots for the CT-index parameter are presented above, the corresponding plots for the remaining performance parameters are compiled in Appendix 4 (see sections A4.3 and A4.4). Table 5 summarizes the calculated maximum deviation envelope limits alongside the aggregate Student's t-test results for all evaluated parameters.

Table 5. Summary of the Student's t-test for mixtures in Task 2.

Parameter	**Max. Deviation (%)**	**Total Number of Sample Sets***	**Number of Sample Sets with Significant Difference (p-value < 0.05)**		
			All t-statistics**	**t-statistic > 0**	**t-statistic < 0**
CT-Index	41.7	13	3 (23.1%)	1	2
RT-Index	49.3	13	8 (61.5%)	3	5
Flexibility	47.7	13	5 (38.5%)	3	2
G_f	16.0	13	3 (23.1%)	1	2
CPR	21.7	13	1 (7.7%)	1	0
Tensile Strength	18.5	13	8 (61.5%)	2	6

*Total number of the corresponding sample sets which were compacted using similar mixture at the same aging level but conditioned using dry and wet methods.

**Percentages were calculated and presented in parenthesis.

As detailed in Table 5, the maximum deviation envelopes across the evaluated parameters ranged from 16.0% to 49.3%. Consistent with the observations from Task 1, parameters representing more fundamental physical measurements, such as G_f and tensile strength, exhibited significantly tighter deviations (16.0% and 18.5%, respectively). In contrast, composite metrics like the RT-index, flexibility, and CT-index exhibited considerably higher maximum deviations, approaching 50%. Additionally, the mixture-specific statistical analysis utilizing the Student's t-tests once again revealed that there is no strong correlation between the deterministic maximum deviation and the actual statistical significance of the conditioning method effect. For instance, while tensile strength exhibited one of the lowest maximum deviations (18.5%), the t-tests identified a statistically significant conditioning method effect in a staggering 61.5% of the evaluated mixtures. Conversely, the RT-index yielded the same rate of statistical significance (61.5%) despite being characterized by a substantially higher maximum deviation of 49.3%.

Furthermore, a critical observation from this summary table is the divergent directionality of the significant results. For every parameter where more than one sample set was significantly affected by the conditioning method, the performance shift was not unidirectional. The concurrent presence of both positive and negative t-statistics across the dataset, even for highly affected parameters like the RT-index and tensile strength, confirms that when the conditioning method does induce a significant change, its directional impact remains mixture-dependent.

To assess the simultaneous impact of the experimental variables across the entire database, a two-way ANOVA was conducted. Table 6 presents the detailed two-way ANOVA output for the CT-index parameter, utilizing the robust HC3 standard error methodology detailed in Appendix 1 (section A1.6).

Table 6. Results of two-way ANOVA based on CT-Index parameter for Task 2.

Parameter	**SSE**	**Degree of Freedom**	**F-statistic**	**p-value**
Conditioning Method	16.10	1	0.232	0.63
Mixture	211,382.14	12	253.797	0.00
Combined Effect	4,317.64	12	5.184	0.00
Residual	9,786.33	141		

An evaluation of the model's main effects reveals two distinct outcomes. First, the main effect of the specific mixture type yielded a highly significant p-value well below the 0.05 threshold ($p = 0.00$). This confirms the expected foundational variance within the database, verifying that the material properties of the different mixtures led to different cracking performance, with average CT-indices ranging from approximately 21 to 180. Conversely, the main effect of the conditioning method was found to be statistically insignificant ($p = 0.63 \gg 0.05$). From a macro-level perspective, this indicates that conditioning the specimens using either the dry or wet method did not induce a systematic, universal shift in the CT-indices across the broad spectrum of evaluated mixtures.

Table 6 also shows that the two-way ANOVA model yielded a statistically significant interaction effect ($p = 0.00$), indicating a non-uniform response. While the conditioning method fails to show a significant overall main effect on the generalized dataset, this significant interaction statistically confirms that the impact of the conditioning method on CT-indices is strictly mixture-dependent.

Finally, the statistical findings from the two-way ANOVA analyses across all evaluated performance metrics and mixtures were aggregated and are summarized in Table 7. A review of these results highlights several distinct behavioral trends regarding how different parameters respond to the experimental variables.

Table 7. Summary of two-way ANOVA results for IDEAL-CT and IDEAL-RT tests in Task 2.

Parameter	Conditioning Method		Mixture Effect		Combined Effect	
	F-statistic	p-value	F-statistic	p-value	F-statistic	p-value
CT-Index	0.232	✗ 0.63	253.797	✓ 0.00	5.184	✓ 0.00
RT-Index	8.547	✓ 0.00	769.524	✓ 0.00	13.849	✓ 0.00
Flexibility	5.213	✓ 0.02	392.396	✓ 0.00	5.402	✓ 0.00
Failure Energy	2.408	✗ 0.12	161.776	✓ 0.00	7.110	✓ 0.00
CPR	0.935	✗ 0.34	27.494	✓ 0.00	1.280	✗ 0.24
Tensile Strength	3.737	✗ 0.06	1,754.051	✓ 0.00	13.002	✓ 0.00

First, the main effect of the mixture type was found to be highly significant across all evaluated parameters, which indicates that the different mixtures in the dataset responded significantly differently under the IDEAL-CT and IDEAL-RT tests. It also confirms that these performance tests are capable of distinguishing among distinct mixtures with varying material properties.

Second, both the RT-index and the flexibility parameters proved to be the most sensitive parameters in the dataset, standing as the only metrics where all three factors were statistically significant. This outcome indicates a dual sensitivity: the choice of wet versus dry conditioning statistically alters the measured RT-index and flexibility across the broader dataset, but the magnitude and direction of that alteration is mixture dependent.

Third, the results showed that only the mixture effect was significant for the CPR parameter, where neither the main effect of the conditioning method nor the combined interaction term reached statistical significance. This indicates that the CPR is fundamentally unaffected by the choice between dry or wet conditioning, either on a macro-level or a mixture-specific basis. Consistent with previous observations, this behavior likely stems from the relatively low signal-to-noise ratio inherent to the CPR metric.

Finally, a distinct and common behavioral trend emerged for the CT-index, G_f, and tensile strength. For these three parameters, the two-way ANOVA yielded an insignificant main effect for the conditioning method (noting that tensile strength was marginally insignificant at $p = 0.06$), but a highly significant combined interaction effect. This indicates a non-uniform response: the conditioning method does not induce a consistent, global shift in these metrics across the board. Instead, the significant interaction indicates that the impact of the conditioning method on these parameters is mixture dependent. This represents a difference for tensile strength compared to the findings in Task 1, where it previously exhibited a uniform global shift regardless of the mixture type.

In a separate analysis conducted for this task, the correlations between mixture water absorption and performance metrics derived from the IDEAL-CT and IDEAL-RT tests were examined. The results revealed no discernible trend for either the IDEAL-CT-related performance metrics or the RT-index from the IDEAL-RT test. Figure 14 shows, as a representative example, the correlation between mixture water absorption and RT-index, showing that no meaningful relationship was observed.

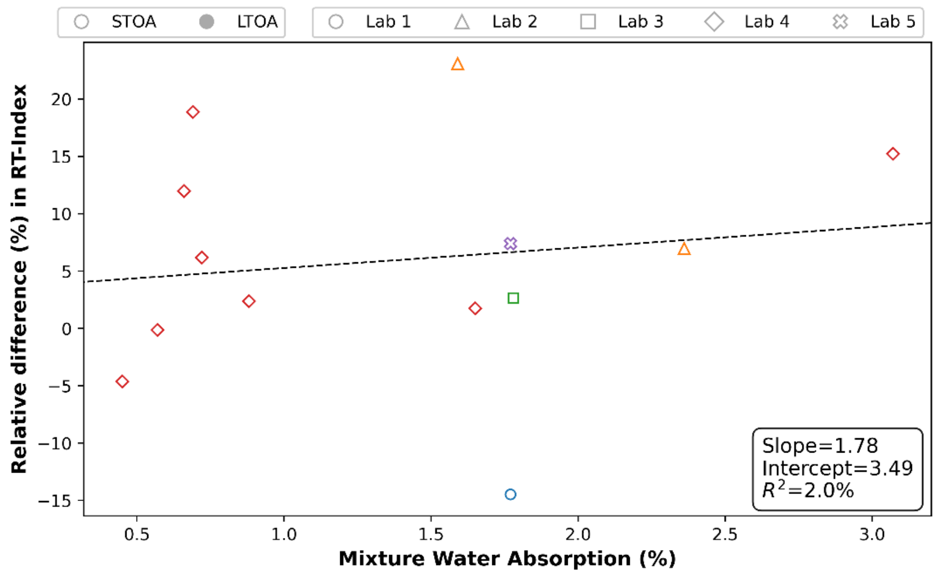


Figure 14. Correlation between the mixture water absorption and relative difference in RT-Index (source: FHWA).

4.3 Task 3: Effect of Heating Protocol

Task 3 was designed to evaluate the specific effect of the loose mixture oven heating protocol on the performance metrics derived from the IDEAL-CT and IDEAL-RT tests. Four experimental oven heating protocols were evaluated in this task, denoted as Methods A through D:

- Method A: the loose mixture is placed in an oven that is already preheated to the target compaction temperature for a duration of exactly 3 hours.
- Method B: the loose mixture is heated in a preheated oven until the internal temperature of the mixture itself reaches the target compaction temperature ± 5.6 ℃ (10 ℉).
- Method C: the loose mixture is placed into a cold oven, which is then turned on and heated until the mixture reaches the target compaction temperature ± 5.6 ℃ (10 ℉).
- Method D: the loose mixture is heated in a preheated oven until it reaches a temperature 27.8 ℃ (50 ℉) below the target compaction temperature.

In this task, ten distinct plant-produced asphalt mixtures were collected and evaluated by four independent laboratories across the United States. As detailed in the supporting volumetric analysis provided in Appendix A5.1, all laboratory-fabricated test specimens were successfully compacted to the target air voids of 7.0% ± 0.5%, with the singular exception of mixture L1-M4, which exhibited air voids systematically lower by approximately 0.5%. Furthermore, all laboratories reported their specific oven heating times. These durations were analyzed as a quality control verification approach to ensure the proper execution of the prescribed heating protocols, as detailed in Appendix A5.2. According to this analysis, while potential procedural discrepancies were identified for four specific mixtures, the intended oven heating protocols were successfully applied across most of the dataset.

To evaluate the specific impact of the heating protocol, this task utilized the IDEAL-RT and IDEAL-CT tests as the primary performance baselines. Accordingly, the RT- and CT-indices, as the primary output parameters of these tests, are plotted for all evaluated mixtures in Figure 15 and Figure 16, respectively.

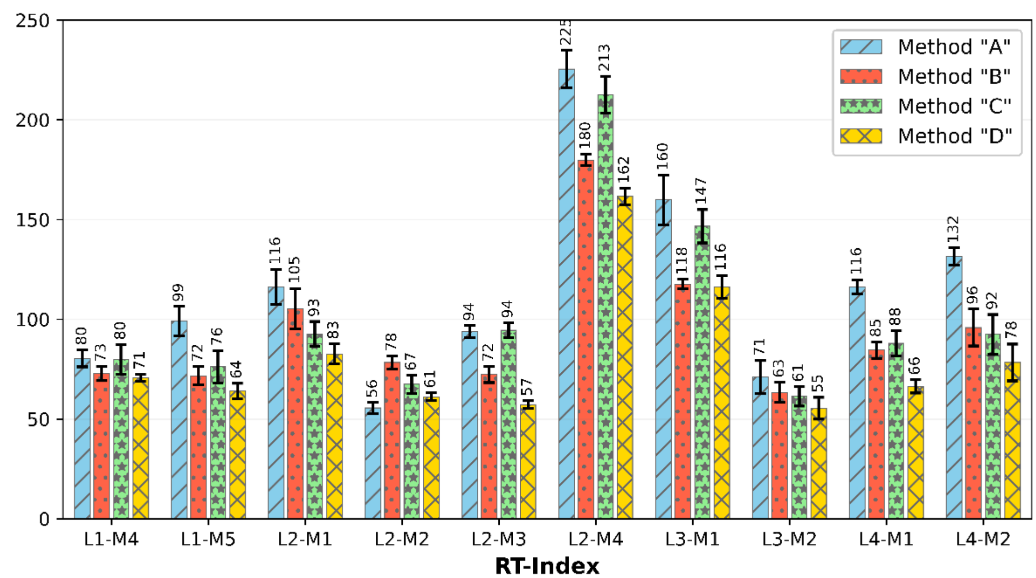


Figure 15. Summary of calculated RT-indices for all evaluated mixtures in Task 3 (source: FHWA).

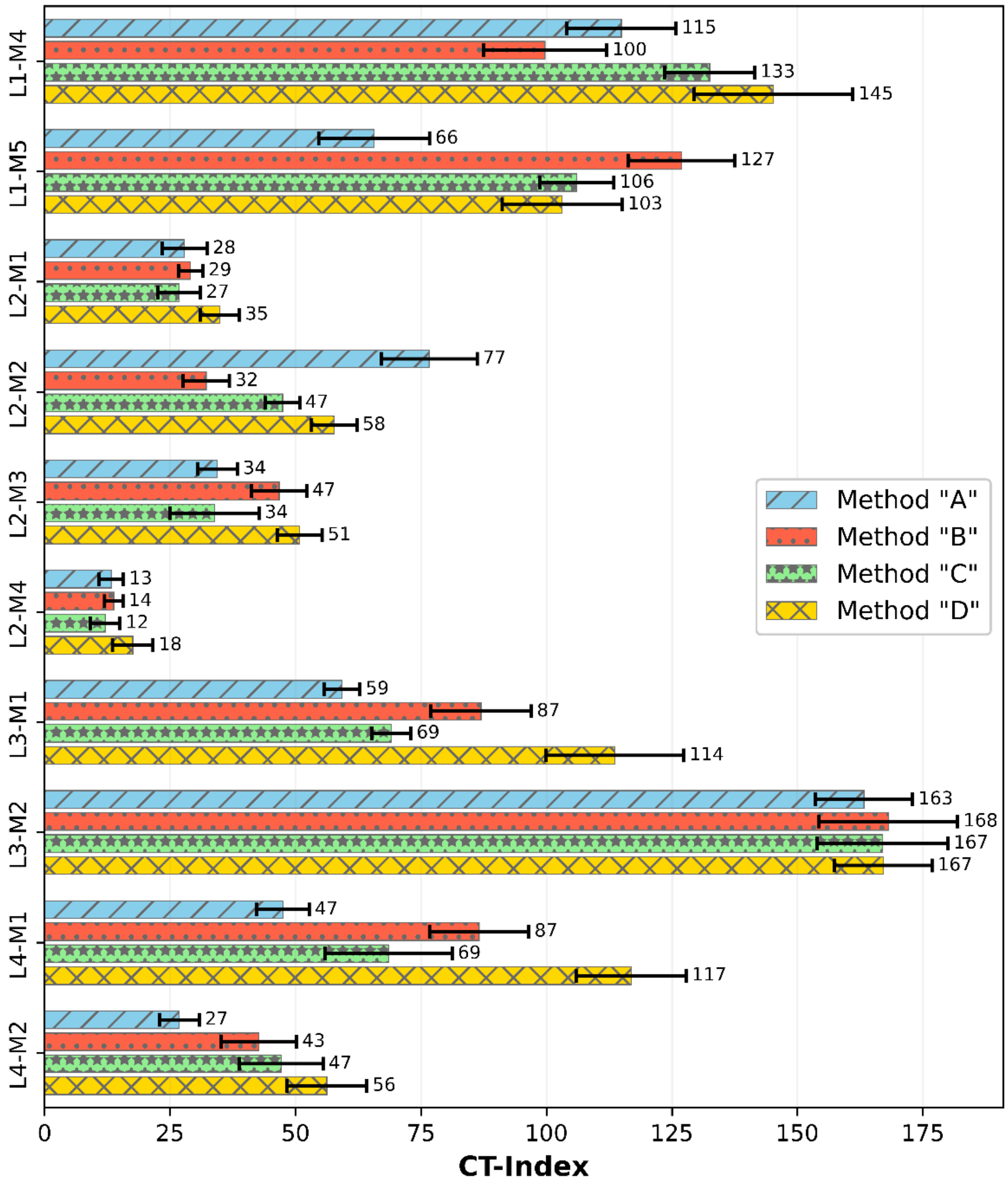


Figure 16. Summary of calculated CT-indices for all evaluated mixtures in Task 3 (source: FHWA).

As illustrated in Figure 16, the IDEAL-CT results indicate that the selected plant-produced mixtures represent a robust set of mixtures. The dataset captures mixtures exhibiting average CT-indices as between 12 and 168. Similarly, the distribution of the IDEAL-RT results in Figure 15 shows a good spread in measured performance, with average RT-indices ranging from 55 to 225. This range across both performance metrics confirms that the assembled database encompasses a broad, representative cross-section of mixture behaviors. Consequently, this helps ensure that the subsequent statistical evaluation of the oven heating protocols' effect is representative of a wide variety of typical mixtures.

Similar to the methodology employed in Tasks 1 and 2, a suite of performance metrics was calculated from the raw testing data files. For the IDEAL-RT, these outputs included the RT-index and tensile strength. For the IDEAL-CT, the extracted parameters encompassed the CT-index, flexibility parameter, G_f, CPR, and tensile strength. Detailed methodologies outlining the calculation of these parameters are documented in Appendix 2 (see page 52).

The effect of the heating protocol was initially visualized utilizing equality plots. Because Task 3 evaluated four distinct oven heating protocols, these visualizations are structured as multi-panel grids. Within each grid, the performance metrics derived from specimens subjected to a specific heating protocol (Methods A, B, C, or D) are plotted against those utilizing Method A, which serves as the experimental reference baseline. Consequently, the first panel simply compares Method A against itself, resulting in all data points resting perfectly on the line of equality. Figure 17 shows this comparative approach using the CT-index parameter as a representative example. As this figure shows, each panel within the plot consists of ten distinct data points, each representing a specific mixture within the dataset, complete with error bars indicating the intra-set standard deviation. The shaded region on the equality plots delineates the maximum deviation envelope, the theoretical details of which were established previously in Section 4.1 (see page 14).

As previously discussed, this envelope serves as a non-statistical, deterministic boundary that defines the maximum percentage difference required to assume parity between the plotted variables. For the CT-index dataset, the bounding envelopes for comparing Methods B, C, and D against the Method A baseline were calculated at 137.8%, 75.7%, and 146.0%, respectively. These exceptionally high bounding envelopes indicate that the choice of oven heating protocol exerts a substantial impact on the resulting CT-indices.

Similar equality plots were generated for the remaining performance parameters and are detailed in Appendix

A5.3. To synthesize these findings, the maximum deviation bounds extracted from these plots are summarized in Table 8. As this table shows, the composite cracking metrics, like CT-index and flexibility parameter, exhibited the highest maximum deviations. These were followed by moderate deviations in the RT-index and tensile strength. Conversely, foundational failure properties like the CPR and G_f exhibited the tightest variance, with G_f yielding the lowest overall maximum deviations (ranging from just 10.1% to 21.4%).

Table 8. Summary of calculated maximum deviation bounds for evaluated performance metrics compared to the Method A baseline.

Parameter	**Maximum Deviation Bound compared to Method A**			
	Method A	**Method B**	**Method C**	**Method D**
CT-index	0.0%	137.8%	75.7%	146.0%
RT-index	0.0%	41.0%	42.5%	75.3%
Flexibility parameter	0.0%	116.0%	68.3%	139.6%
G_f	0.0%	10.1%	19.8%	21.4%
CPR	0.0%	15.5%	21.5%	18.2%
Tensile strength	0.0%	26.4%	23.8%	30.6%

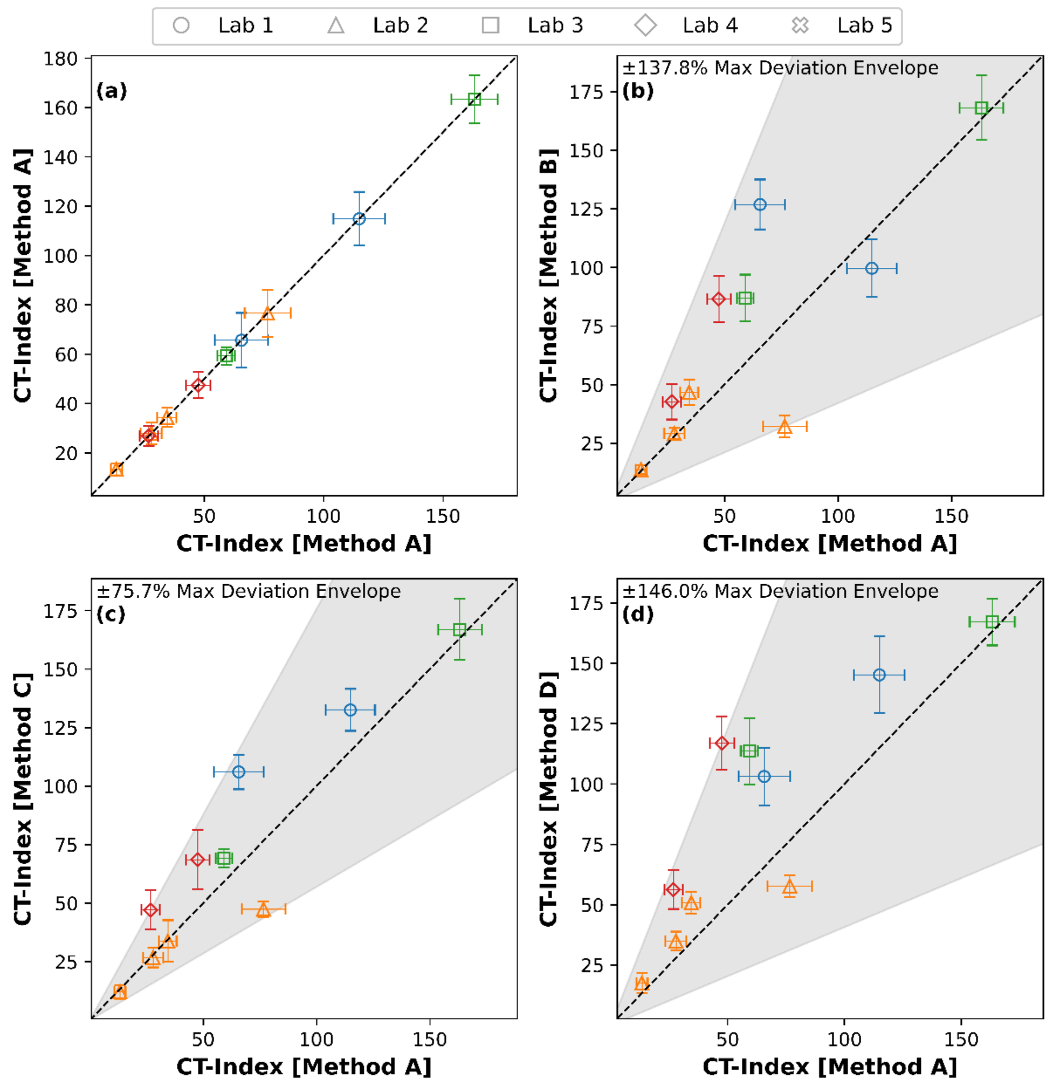


Figure 17. Equality plots comparing the CT-indices of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

An additional critical observation from Figure 17 is that one specific mixture from Laboratory 2 (L2-M2) exhibited a reversed behavioral response compared to the general trend of the broader dataset. Specifically, Figure 17(b) shows that while the CT-indices derived using Method B are generally higher than those from Method A (placing most data points above the line of equality), L2-M2 mixture significantly disobeys this trend. The average CT-index for L2-M2 under Method A is 77, which is substantially higher than its average of 32 under Method B, placing this specific data point well below the line of equality. Similar reversed behavior was documented for this mixture when comparing Methods C and D against Method A. Furthermore, a review of the equality plots for the remaining performance metrics (detailed comprehensively in Appendix

A5.3) reveals that this exact same anomalous behavior for mixture L2-M2 is also present in the RT-index and flexibility parameter results. Interestingly, this mixture had the highest aggregate water absorption at 2.36%, while all other mixtures evaluated for this task ranged between 0.69% and 2.07%.

However, high absorption alone may not explain the behavior, as L1-M5 also had a high absorption rate (2.07%) but did not exhibit this reverse trend.

In addition to the deterministic maximum deviation analysis, a statistical evaluation of the four oven heating protocols was conducted on an individual, per-mixture basis utilizing a one-way ANOVA, with additional details provided in Appendix A5.4. A summary of these results reveals that parameters such as the RT-index, tensile strength, CT-index, flexibility parameter, and G_f were highly sensitive to the heating protocol, yielding statistically significant differences across most evaluated mixtures. Conversely, the CPR was the only metric that remained largely unaffected by the heating variations.

However, because the one-way ANOVA serves strictly as an omnibus statistical test, indicating only that a difference exists among the evaluated methods without identifying the specific contrasting pairs, a Games-Howell post-hoc analysis was subsequently performed. This specific post-hoc method was selected to appropriately account for the inhomogeneity of variances within the dataset, and the detailed pairwise outputs are compiled in Appendix A5.5. A comprehensive review of these pairwise comparisons revealed that no two specific heating methods consistently produced the most significant or insignificant differences across the board. Instead, the results yielded a highly fluctuating mix of significant pairs depending on the material. Ultimately, this shows that while the choice of loose mixture oven heating protocol unequivocally impacts the measured performance metrics, this effect is likely mixture dependent. Other variables that might have affected the conclusiveness of this evaluation include sample variability, equipment differences, and interlaboratory inconsistencies.

To assess the simultaneous impact of the experimental variables across the entire database, a two-way ANOVA was conducted. Table 9 presents the detailed two-way ANOVA output for the CT-index parameter, utilizing the robust HC3 standard error methodology detailed in Appendix 1 (Section A1.6) to properly account for the dataset's variance characteristics.

Table 9. Results of two-way ANOVA based on CT-Index parameter for Task 3.

Parameter	**SSE**	**Degree of Freedom**	**F-statistic**	**p-value**
Conditioning Method	15,509.68	3	74.441	0.00
Mixture	497,886.23	9	796.563	0.00
Combined Effect	34,554.95	27	18.428	0.00
Residual	14,792.69	213		

Analyzing the main effects within the model highlights two key findings. First, the main effect of the mixture type is highly significant ($F = 796.563, p = 0.00$). This serves as a critical baseline validation of the dataset's diversity, proving that distinct material and volumetric properties drive substantial differences in cracking resistance and successfully capturing a broad CT-index range from approximately 12 to 168. Second, the oven heating protocol itself emerged as a highly significant main effect ($F = 74.441, p = 0.00$). This demonstrates that, on a generalized level, altering the loose mixture heating procedure systematically alters the measured CT-index across the broader dataset.

The ANOVA output also reveals a highly significant interaction between the heating protocol and the mixture type ($p = 0.00$). This indicates that the laboratory specimens do not respond uniformly to procedural changes. Therefore, while the heating method does exert a significant main effect, the exact nature of that impact, specifically its magnitude and direction, is likely dictated by the unique characteristics of each individual asphalt mixture.

Finally, the statistical findings from the two-way ANOVA analyses across all evaluated performance metrics and mixtures were aggregated and are summarized in Table 10. A review of these results highlights a uniform behavioral trend regarding how the different parameters respond to the experimental variables in this specific task.

Table 10. Summary of two-way ANOVA results for IDEAL-CT and IDEAL-RT tests in Task 3.

Parameter	**Conditioning Method**		**Mixture Effect**		**Combined Effect**	
	F-statistic	**p-value**	**F-statistic**	**p-value**	**F-statistic**	**p-value**
CT-Index	74.441	✓ 0.00	796.563	✓ 0.00	18.428	✓ 0.00
RT-Index	293.703	✓ 0.00	1058.842	✓ 0.00	34.617	✓ 0.00
Flexibility	122.917	✓ 0.00	960.550	✓ 0.00	20.731	✓ 0.00
Failure Energy	13.922	✓ 0.00	693.704	✓ 0.00	8.028	✓ 0.00
CPR	13.424	✓ 0.00	23.169	✓ 0.00	2.490	✓ 0.00
Tensile Strength	289.279	✓ 0.00	1484.432	✓ 0.00	28.877	✓ 0.00

Unlike the varying parameter sensitivities observed in the previous tasks, Table 10 reveals universal statistical significance. For every single evaluated parameter—including CT-index, RT-index, flexibility parameter, CPR, G_f—all three statistical effects (the mixture main effect, the heating protocol main effect, and the combined interaction effect) yielded p-values well below the 0.05 threshold (all reported as $p = 0.00$).

From an engineering perspective, this total uniformity dictates three critical conclusions regarding the loose mixture oven heating process. First, it reaffirms that the tests accurately capture the differences across the broad set mixtures. Second, it provides strong evidence that the choice of heating protocol fundamentally and systematically alters the baseline results for every measured performance metric; no parameter is immune to the procedural variations between Methods A, B, C, and D. Lastly, the universally significant combined interaction effect shows that this impact is never uniform. Across all metrics, the specific magnitude, severity, and directional shift caused by the heating protocol remain entirely dependent on the unique material and volumetric characteristics of the individual asphalt mixture.

4.4 Task 4: Effect of Trimming Method

Task 4 was specifically designed to investigate the effect of laboratory specimen trimming methods on the performance metrics derived from HWTT tests. To assess this, three distinct specimen fabrication methods were evaluated, denoted as Methods A through C:

- Method A (direct compaction): specimens were compacted directly to the final testing height of 62 mm, requiring no saw cutting.
- Method B (single-side trimming): specimens were initially compacted to a height of 115 mm and subsequently saw-cut on one face to achieve the 62 mm target height. During testing, the uncut (molded) surface was oriented upwards to interface directly with the HWTT wheel.
- Method C (double-side trimming): specimens were initially compacted to 115 mm and then trimmed on both the top and bottom faces to yield the final 62 mm testing height.

For this experimental phase, seven distinct plant-produced asphalt mixtures were collected and evaluated across five independent testing laboratories in the United States. As documented in the volumetric analysis in Appendix A6.1, the laboratories successfully compacted all test specimens to the designated target air void content of 7.0% ± 0.5%. Furthermore, a review of this volumetric data showed that for this dataset the choice of trimming method did not introduce any significant bias or unintended variation into the final air void contents of the tested specimens.

To evaluate the specific impact of the specimen trimming method on HWTT performance, the raw rutting curves, with the maximum rut depths shown in Figure 18, were analyzed utilizing the two-part power (2PP) model via the AutoHWTT software (Abdollahi *et al.* TBD). A comprehensive background on this analytical methodology is documented in Appendix A2.3. From this analysis, a suite of performance parameters was extracted for each tested specimen. The Corrected Rut Depth at 20,000 passes (CRD20) and the Stripping Number (SN)—representing viscoplastic deformation and moisture susceptibility, respectively—are plotted for all evaluated mixtures in Figure 19 and Figure 20. It should be noted that the data presented in this figure includes all four experimental replicates for each sample set; data points flagged by the automated outlier detection algorithm were intentionally retained for this macro-level visualization. Similar plots for the remaining performance parameters, both inclusive and exclusive of identified outliers, are detailed comprehensively in Appendix A6.2.

A visual assessment of the CRD20 results in Figure 19 reveals that across most of the evaluated mixtures, specimens prepared using Method B (single-side trimming) exhibited higher viscoplastic deformation compared to their counterparts. Conversely, specimens prepared using Method C (double-side trimming) consistently yielded the lowest or comparable CRD20 values relative to the direct compaction baseline of Method A. While these visual trends can be interpreted that trimming protocols impact shear resistance, statistical analysis is required to confirm these observations with a higher degree of confidence.

Similarly, a visual assessment of the SN data in Figure 20 indicates a broad spectrum of moisture susceptibility, with average SN values ranging from 1,600 to 19,000 passes. This extensive range confirms that the selected dataset provides a diverse foundation for evaluating stripping performance. Furthermore, for most of the evaluated mixtures (e.g., L2-M1, L2-M2, L3-M1, L4-M2, and L5-M1), the average SN values of the specimens prepared using Method C were lower than those prepared using Method A. Mechanistically, this can be likely attributed to the double-sided saw-cutting process; cutting exposes the internal aggregate structure by shearing through the protective binder and mastic films, thereby providing more pathways for moisture infiltration and accelerating the stripping potential during HWTT loading. However, it must be noted that this behavior was not universally observed across the dataset. Mixture L1-M4, in particular, exhibited a reversed trend, maintaining higher SN values for the trimmed specimens even when statistical outliers were excluded from the analysis (as detailed in Figure A6.7).

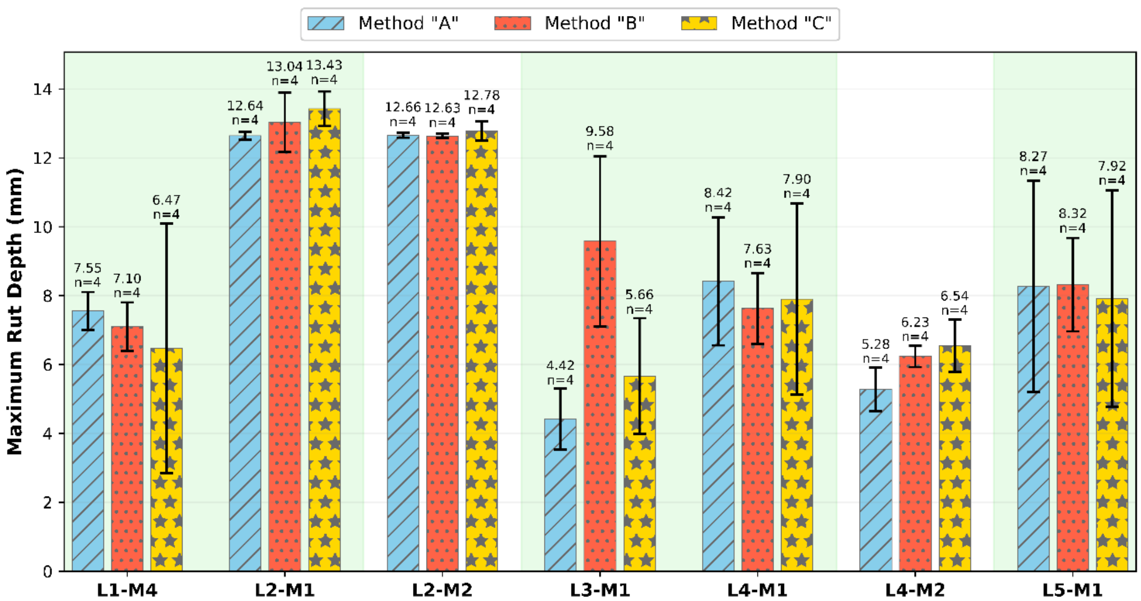


Figure 18. Summary of maximum rut depth for evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

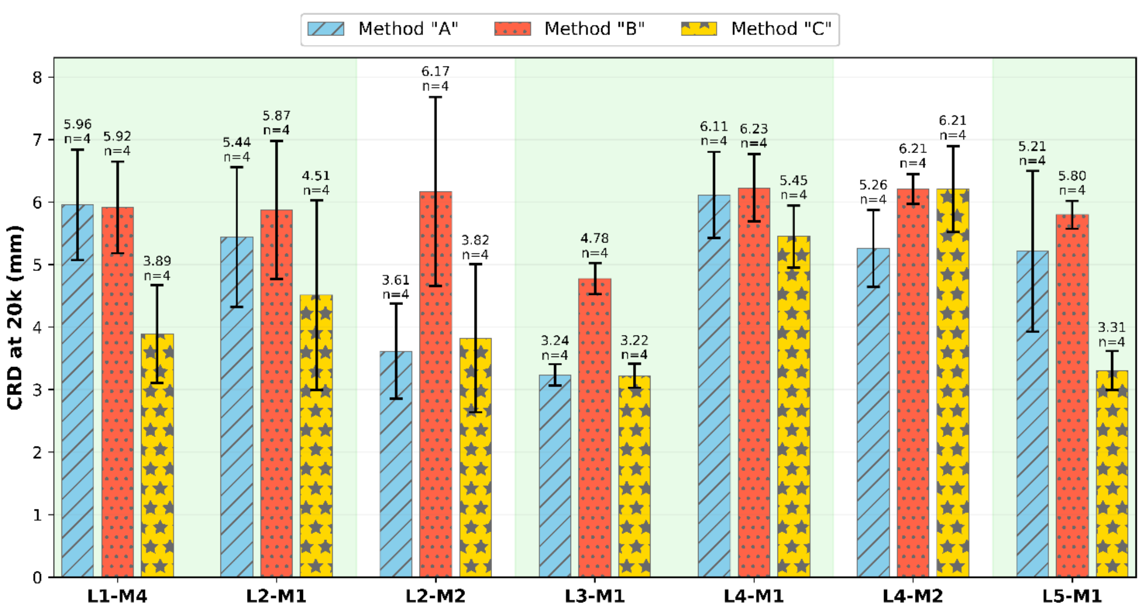


Figure 19. Summary of CRD20 for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

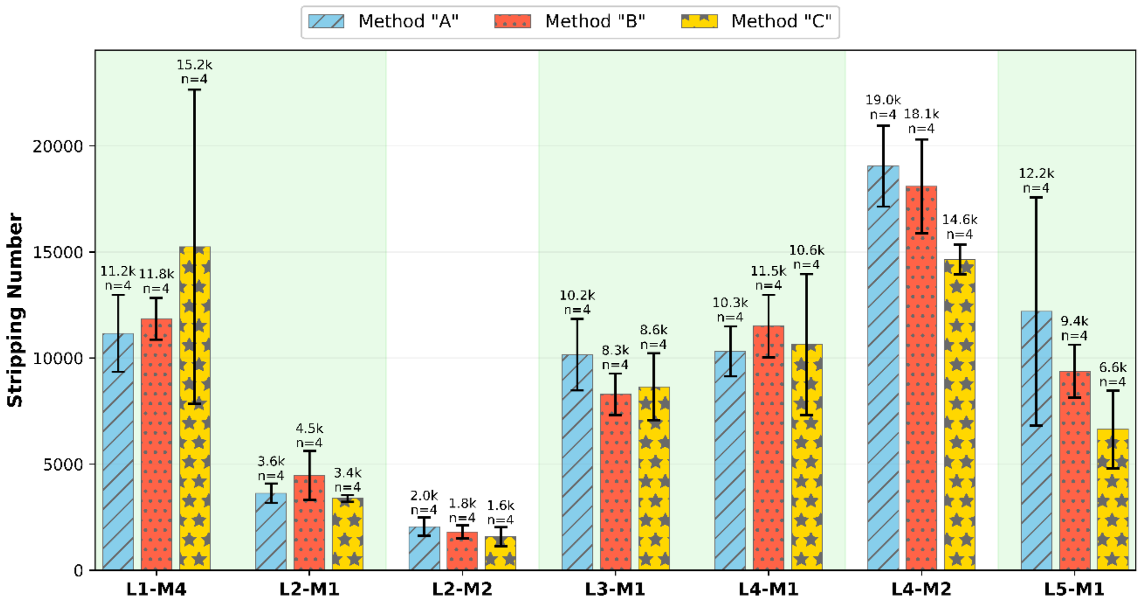


Figure 20. Summary of SN for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

The effect of the trimming method was initially visualized utilizing equality plots. Because Task 4 evaluated three distinct trimming procedures, these visualizations are structured as multi-panel grids cross comparing the methods (specifically, Method A vs. Method B, Method A vs. Method C, and Method B vs. Method C). Figure 21 shows this comparative approach using the CRD20 parameter, inclusive of all replicates, as a representative example. As depicted, each panel within the plot contains seven distinct data points—each representing a specific mixture within the dataset—complete with error bars indicating the intra-set standard deviation. The shaded region on the equality plots delineates the maximum deviation envelope, the theoretical details of which were established previously in Section 4.1 (see page 14).

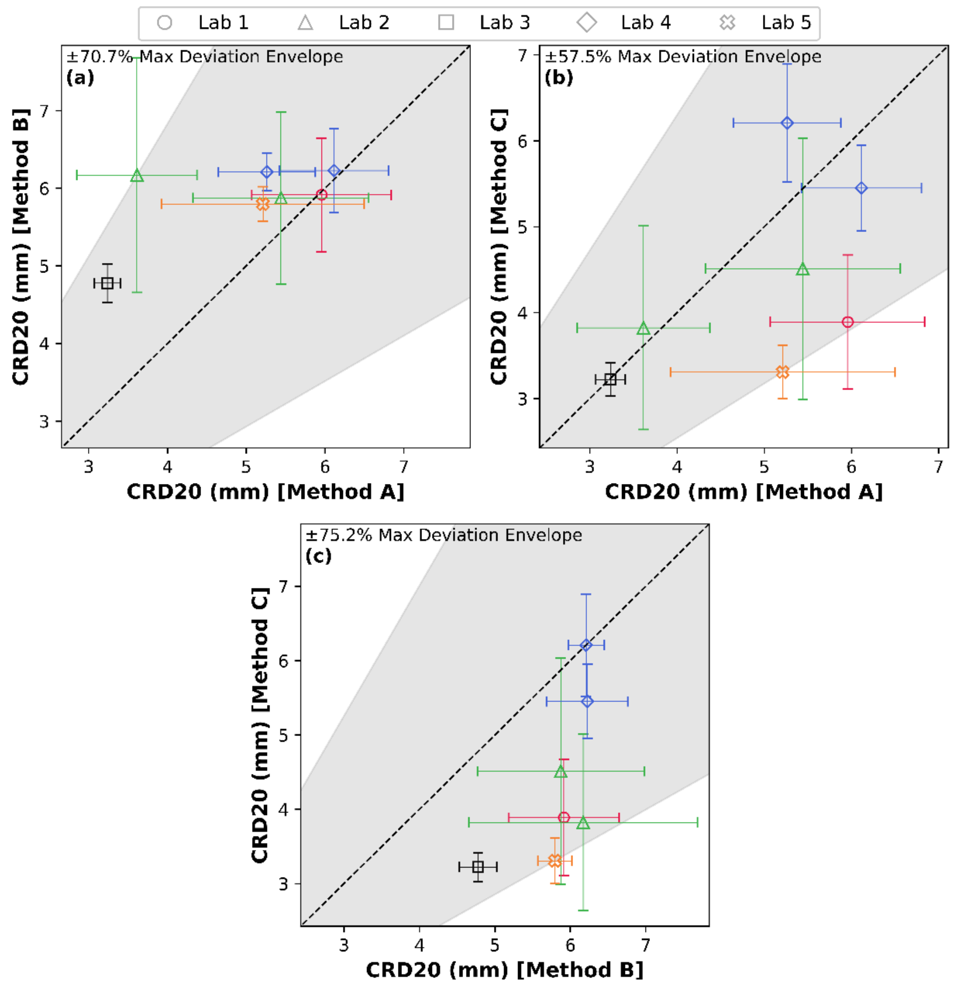


Figure 21. Equality plots comparing CRD20 of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

As previously discussed, this envelope serves as a non-statistical, deterministic boundary that defines the maximum percentage difference required to assume parity between the plotted variables. For the CRD20 dataset, the bounding envelopes for comparing Methods A vs. B, A vs. C, and B vs. C were calculated at 70.7%, 57.5%, and 75.2%, respectively. These exceptionally high bounding envelopes indicate that the choice of trimming method exerts a substantial impact on the resulting CRD20 values.

Similar equality plots were generated for the remaining performance parameters (TRD, SN, SIP, and Adjusted SIP), both inclusive and exclusive of statistical outliers, and are detailed comprehensively in Appendix A6.3. To synthesize these findings, the maximum deviation bounds extracted from all plots are summarized in Table 11.

Table 11. Summary of calculated maximum deviation bounds on effect of trimming method for evaluated performance metrics based on all replicates.

Parameter	Maximum Deviation Bound		
	Method A vs. Method B	**Method A vs. Method C**	**Method B vs. Method C**
TRD	117.0%	28.3%	69.1%
CRD20	70.7%	57.5%	75.2%
SN	30.0%	83.7%	41.3%
SIP	47.2%	28.1%	14.9%
Adjusted SIP	69.4%	143.7%	109.1%

A close examination of Table 11 reveals a divergence in behavior between viscoplastic deformation and moisture susceptibility metrics. Logically, one might assume that the maximum deviation between Methods A and C (uncut vs. double-cut) would consistently exceed the deviation between Methods A and B (uncut vs. single-cut), simply because Method C alters two specimen faces instead of one. However, this assumption only holds true for two moisture-driven parameters: SN and adjusted SIP. For these specific metrics, exposing two internal aggregate faces to the water bath predictably causes a more severe deviation from the uncut baseline than exposing just one.

Conversely, an exact opposite trend was observed for other parameters (CRD20, TRD, and SIP): the maximum deviation for Method A vs. Method B is substantially higher than Method A vs. Method C (e.g., reaching an extreme 117.0% vs. 28.3% for TRD). This can be attributed to the hypothesis that double-sided trimming (Method C) likely produces a more structurally symmetric and volumetrically balanced specimen for mechanical shear loading, yielding results closer to the uniform direct-compaction specimens (Method A). Single-sided trimming (Method B), however, introduces a volumetric and density asymmetry that destabilizes viscoplastic behavior. Ultimately, it is noted that while this deterministic maximum deviation bounding analysis highlights the worst-case mixture variations, a rigorous statistical analysis is required to determine the significance of these shifts across the broader dataset.

In the next step, a one-way ANOVA was performed on a per-mixture basis to statistically assess the broader impact of the three trimming methods on the HWTT output parameters. The details of this analysis are provided in Appendix A6.4, with a summary of the results presented in Table 12. It is important to restate that a statistically significant one-way ANOVA ($p \leq 0.05$) confirms that a fundamental difference exists among the evaluated trimming methods for that specific mixture, but as an omnibus test, it does not identify which specific methods differ from one another.

Table 12. Summary of one-way ANOVA analysis on effect of trimming method.

Parameter	Total Number of Mixtures	Number of mixtures with $p \leq 0.05$	
		All replicates	**Excluding Outliers**
TRD	7	2 (28.6%)	4 (57.1%)
CRD20	7	4 (57.1%)	5 (71.4%)
SN	7	1 (14.3%)	2 (28.6%)
SIP	7	2 (28.6%)	5 (71.4%)
Adjusted SIP	7	2 (28.6%)	3 (42.9%)

A review of the summarized data in Table 12 highlights several critical insights about trimming methods. First, a considerable contrast exists between the statistical results and the deterministic maximum deviations. Based on the dataset inclusive of all replicates, only one and two mixtures (out of the seven evaluated) exhibited a statistically significant difference in their SN and Adjusted SIP values, respectively. In contrast, the earlier deterministic maximum deviation analysis flagged these exact parameters as being heavily altered by double-sided trimming (Method C), showing extreme maximum deviations of 83.7% and 143.7% compared to direct compaction. This observation perfectly illustrates the fundamental difference between deterministic and statistical analyses: the maximum deviation metric captures an isolated, worst-case scenario within a dataset of mixtures, whereas the one-way ANOVA demonstrates that this extreme sensitivity is more of an exception rather than the rule.

Another critical observation from Table 12 is the profound impact of statistical outliers on small datasets. As detailed in Appendix A1.1.2, each sample set contains a maximum of four HWTT replicates, and the automated outlier detection algorithm conservatively removes no more than one anomalous data point. While state agencies and practitioners may practically hesitate to discard any data from such small sample sizes, the table clearly demonstrates that isolating just one outlier significantly cleans the dataset by reducing internal variance. For every single evaluated parameter, excluding these outliers increased the number of mixtures that triggered statistical significancy ($p \leq 0.05$).

Finally, a Games-Howell post-hoc analysis was conducted to evaluate the pairwise significance between the trimming methods for each individual mixture (detailed in Appendix A6.5). A review of these results revealed some discrepancies where the omnibus one-way ANOVA indicated a significant dataset variance, yet the post-hoc analysis failed to identify any significant pairwise combinations. For example, evaluating the SIP parameter for mixture L5-M1 (using filtered replicates in Table A6.1) yielded a significant ANOVA result ($p = 0.01; F = 10.744$), but the lowest pairwise Games-Howell p-value was only 0.093 (comparing Methods A and C). These discrepancies are directly attributable to the highly conservative nature of the Games-Howell adjustment, which imposes a severe penalty on the degrees of freedom when dealing with small sample sizes and unequal variances. Consequently, it can be deduced that the initial standard one-way ANOVA—which assumes normality and homogeneity—likely produced a Type I error (false positive) for these specific, highly variable edge cases.

To assess the simultaneous impact of the experimental variables across the entire database, a two-way ANOVA was conducted. Table 13 presents the detailed two-way ANOVA output for the CRD20 parameter (utilizing the dataset inclusive of all replicates). This analysis employs the robust HC3 standard error methodology, as detailed in Appendix A1.6, to properly account for the dataset's specific variance and heteroskedasticity characteristics.

Table 13. Results of two-way ANOVA based on CRD20 parameter for Task 4 using all replicates.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	139.85	2	99.032	0.00
Mixture	152.23	6	35.933	0.00
Combined Effect	37.39	12	4.413	0.00
Residual	44.48	63		

Analyzing the main effects within the model highlights two key findings. First, the inherent influence of the specific mixture type yielded a highly significant result ($F = 35.933, p = 0.00$). This serves as a critical baseline validation of the dataset's overall diversity, indicating that the distinct material and volumetric properties of the evaluated mixtures have different viscoplastic deformation behaviors. Second, the specimen trimming method itself emerged as a highly significant main effect ($F = 99.032, p = 0.00$). From a macro-level perspective, this observation demonstrates that altering the specimen preparation and trimming procedure systematically shifts the measured CRD20 values across the broader dataset.

Additionally, the two-way ANOVA output also reveals a statistically significant combined interaction effect between the trimming method and the mixture type ($F = 4.413, p = 0.00$). This indicates a non-uniform response among the laboratory specimens to the procedural changes. Therefore, while the choice of trimming method exerts a highly significant global influence on viscoplastic deformation measurements, the exact nature of that impact—specifically its magnitude and direction—is dictated by the unique material characteristics of each individual asphalt mixture.

To conclude the statistical evaluation, the two-way ANOVA outputs across all evaluated mixtures and performance metrics were aggregated, as summarized in Table 14. An examination of these global results uncovers a consistent behavioral pattern regarding how the various HWTT parameters react to the experimental variables introduced in this task.

Table 14. Summary of two-way ANOVA results for HWTT parameters in Task 4.

Parameter	**Conditioning Method**		**Mixture Effect**		**Combined Effect**	
	F-statistic	**p-value**	**F-statistic**	**p-value**	**F-statistic**	**p-value**
Based on all replicates (including outliers)						
TRD	2.452	✗ 0.09	444.737	✓ 0.00	1.947	✓ 0.04
CRD20	99.032	✓ 0.00	35.933	✓ 0.00	4.413	✓ 0.00
SN	3.954	✓ 0.02	317.292	✓ 0.00	2.225	✓ 0.02
SIP	6.412	✓ 0.00	438.417	✓ 0.00	2.119	✓ 0.03
Adjusted SIP	0.488	✗ 0.61	286.786	✓ 0.00	2.799	✓ 0.00
Based on filtered replicates (excluding outliers)						
TRD	0.185	✗ 0.83	647.556	✓ 0.00	5.091	✓ 0.00
CRD20	140.834	✓ 0.00	69.022	✓ 0.00	9.835	✓ 0.00
SN	4.564	✓ 0.02	274.001	✓ 0.00	4.062	✓ 0.00
SIP	3.111	✗ 0.05	839.415	✓ 0.00	4.957	✓ 0.00
Adjusted SIP	0.374	✗ 0.69	586.025	✓ 0.00	3.176	✓ 0.00

An examination of the two-way ANOVA results in Table 14 reveals several insights. First, regardless of whether outliers are included or excluded, the main effect of the mixture type remains universally significant ($p = 0.00$) across all parameters. This observation is an ideal outcome, as it validates that the selected metrics—TRD, CRD20, SN, SIP, and Adjusted SIP—are capable of distinguishing between varying material properties and volumetric characteristics. Furthermore, it shows that the selected mixtures provide a highly robust, comprehensive, and diverse dataset, forming a reliable baseline for advanced statistical evaluation.

Second, analyzing the main effect of the specimen trimming method suggests a mixed behavior among the parameters. For the TRD and the Adjusted SIP, the trimming method does not exert a statistically significant global impact, maintaining high p-values regardless of outlier filtration. Conversely, the trimming method consistently and significantly alters both the viscoplastic deformation (CRD20) and the SN across both datasets. This indicates that saw-cutting the aggregate structure changes the pure shear resistance and the onset of moisture damage. In the case of the conventional SIP, the results presented a transitional behavior: it registered as statistically significant when all replicates were analyzed, but the main effect of the trimming method became borderline insignificant (p = 0.05) once the statistical outliers were excluded.

Finally, an assessment of the combined interaction effect reveals a universal trend: regardless of the specific parameter or the inclusion of outliers, the interaction between the trimming method and the mixture type is always highly statistically significant ($p < 0.04$). Mechanistically, this means that laboratory specimens do not respond uniformly to the saw-cutting process. Even if the trimming method does not produce a significant global shift for a metric like TRD or Adjusted SIP, the exact nature of the trimming impacts remains dependent on the material characteristics of the individual asphalt mixture. Ultimately, this significant interaction proves that there is no universal "correction factor" that can be applied to trimmed HWTT specimens, as the physical response to cutting is mixture dependent.

5 Summary and Conclusion

In an attempt to understand inter-laboratory discrepancies often observed in BMD testing, this study investigated the impact of several common specimen fabrication variables on mixture performance. A robust experimental program was executed across five independent laboratories utilizing 17 distinct plant-produced asphalt mixtures, which provided a diverse spectrum of binder grades, aggregate absorptions, and recycled material contents. The evaluation was divided into four primary tasks assessing variables in G_{mm} calculations, specimen conditioning, loose mixture oven heating protocols, and HWTT specimen trimming. Over 1,100 discrete data points were generated and rigorously analyzed utilizing IDEAL-CT, IDEAL-RT, and HWTT testing results, supported by advanced statistical models including robust HC3 two-way ANOVA and Games-Howell post-hoc testing. The findings of this study can be summarized for each task as listed below.

- Task 1 evaluated whether using STOA versus LTOA G_{mm} values significantly alters the target air voids and the resulting IDEAL-CT and IDEAL-RT performance metrics. The main findings of this task are as follows:
    - From a practical perspective, this study showed that using the STOA G_{mm} for all testing is adequate and will produce statistically similar air void and performance test results to using a more rigorous approach, however, it would be conservative to target the upper half of the acceptable air void range (e.g., 7.0% to 7.5%) to ensures that the true specimen air voids do not inadvertently drop below the minimum threshold due to the reference G_{mm} underestimation.
    - Evaluating specimens using LTOA versus STOA G_{mm} targets produced only marginal density shifts, translating to a maximum discrepancy of just 0.5% in the calculated specimen air voids.
    - Student's t-test analyses revealed that for most of the evaluated mixtures, the choice of G_{mm} aging level did not cause a statistically significant shift in their baseline cracking performance metrics, such as the CT-index.
    - The two-way ANOVA results showed that the effect of G_{mm} aging level on RT-index was significant, however the observed shift was mixture dependent and did not consistently increase or decrease the RT-Index in the same direction.
- Task 2 evaluated the effect of pre-test specimen conditioning method on the performance metrics from IDEAL-CT and IDEAL-RT tests. The main findings of this task are as follows:

    - From a practical perspective, this study showed that water- and air-conditioning methods produce statistically similar results for IDEAL-CT testing, whereas they may yield different results for IDEAL-RT testing. It should be noted that this sensitivity to the conditioning method in the IDEAL-RT results can be partially attributed to the inherently lower testing variability associated with this metric, which allows statistical models to more readily flag procedural differences as significant.
    - The RT-index and flexibility parameters were more sensitive to the conditioning method, however the observed shift was mixture dependent and did not consistently increase or decrease the testing in the same direction.
    - The effect of specimen conditioning method on the CT-index was not significant, the Student's t-test results showed that for 10 out of the 13 mixtures evaluated, the conditioning method had a statistically insignificant impact.
- Task 3 evaluated how the material thermal history, specifically the oven heating protocol, affects the performance metrics from IDEAL-CT and IDEAL-RT tests. The findings from this task can be summarized as follows:
    - From a practical perspective, this study demonstrated that a consistent oven heating protocol is essential when performing BMD testing.
    - The choice of oven heating protocol exerts a substantial impact on measured performance metrics, though the sensitivity varies by parameter. Composite metrics like the CT-index and flexibility parameter were highly sensitive, exhibiting extreme maximum deviations of up to 146%, whereas more fundamental properties like failure energy exhibited tighter variance. Additionally, post-hoc pairwise statistical comparisons revealed that no two specific heating methods consistently produced the most significant differences across the board, as the effects fluctuated depending on the mixture.
    - The two-way ANOVA revealed that the oven heating protocol had a statistically significant main effect on every single evaluated parameter, proving that procedural changes fundamentally alter the baseline results universally. Furthermore, the analysis yielded a highly significant combined interaction effect for all metrics, indicating that the magnitude and direction of these effects are mixture dependent.
- Task 4 evaluated how the trimming method used for preparing HWTT specimens effect the rutting and stripping performance test results. The main findings of this task can be summarized as follows:
    - From a practical perspective, this study demonstrated that the specimen trimming method should be consistent when conducting HWTT testing. Because trimming alters both rutting and moisture susceptibility differently depending on the mixture, comparing performance results across specimens prepared with different trimming methods will lead to inaccurate conclusions.
    - The choice of trimming method did not introduce any systematic bias into the final specimen air voids, and all specimens were successfully compacted within the required range.

- Trimming methods significantly altered both viscoplastic deformation (CRD20) and moisture susceptibility (SN). Double-sided trimming exposes more internal aggregate faces, which likely increases water infiltration and accelerates moisture damage. On the other hand, single-sided trimming likely introduces a volumetric asymmetry that destabilizes viscoplastic behavior, reducing pure shear resistance compared to directly compacted specimens.
- The two-way ANOVA revealed that the trimming method has a highly significant effect on CRD20 and SN, and to a slightly lesser extent, SIP. However, consistent with other tasks, the model showed a highly significant interaction between the trimming method and mixture type across all parameters, indicating this effect is mixture-dependent.

References

AASHTO, 2011. Determining the Creep Compliance and Strength of Hot Mix Asphalt (HMA) Using Indirect Tensile Test Device.

AASHTO, 2023. Standard Method of Test for Hamburg Wheel-Track Testing of Compacted Asphalt Mixtures.

Abdollahi, S.F., Jahangiri, B., Leavitt, A., and Arras, B., TBD. Two-Part Power (2PP) Model to Evaluate the Rutting and Moisture Susceptibility of Asphalt Mixtures Using Hamburg Wheel Tracking Test. *Transportation Research Record*.

ASTM International, 2019. Standard Test Method for Determination of Cracking Tolerance Index of Asphalt Mixture Using the Indirect Tensile Cracking Test at Intermediate Temperature.

ASTM International, 2022. Standard Test Method for Determination of Rutting Tolerance Index of Asphalt Mixture Using the Ideal Rutting Test.

Boz, I., Coffey, G.P., Habbouche, J., Diefenderfer, S.D., and Ozbulut, O.E., 2022. A Critical Review of Monotonic Loading Tests to Evaluate Rutting Potential of Asphalt Mixtures. *Construction and Building Materials*, 335, 127484.

Boz, I., Habbouche, J., Taylor, A., and Moore, N., 2025. *Ruggedness Study of Specimen Preparation and Fine-Tuning of Test Methods for IDT-CT and IDT-HT Test*. Charlottesville, VA: Virginia Transportation Research Council, No. VTRC 26-R22.

Cooley Jr, L.A. and Williams, K., 2013. Aggregate absorption in HMA mixtures.

Elwardany, M., Veginati, V., Andriescu, A., and Mensching, D., 2024. Understanding the Sensitivity of Mixture Cracking Performance Tests to Binder Properties and Long-Term Aging. *Transportation Research Record*, 2678 (1), 335–349.

Games, P.A. and Howell, J.F., 1976. Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study. *Journal of Educational Statistics*, 1 (2), 113–125.

Habbouche, J., Boz, I., and Diefenderfer, S.D., 2022. *Interlaboratory Study for the Indirect Tensile Cracking Test at Intermediate Temperature: Phase II*. Charlottesville, VA: Virginia Transportation Research Council, No. VTRC 23-R3.

Habbouche, J., Boz, I., Diefenderfer, S.D., and Bilgic, Y.K., 2021. *Round Robin Testing Program for the Indirect Tensile Cracking Test at Intermediate Temperature: Phase I*. Charlottesville, VA: Virginia Transportation Research Council, No. VTRC 22-R3.

Hajj, E.Y., Aschenbrener, T., and Nener-Plante, D., 2022. Examples of Successful Practices with State Implementation of Balanced Design of Asphalt Mixtures. *Transportation Research Record*, 2676 (5), 44–66.

Hajj, E.Y., Elias, N.G., Khanal, B., Alrajhi, A., Hand, A., Bennert, T., Zhou, F., Epps, J., Chan, K.M., Decker, C.S., Patel, H., and Duval, J., 2025. *Balanced Mix Design: Rutting Performance Test*. Greenbelt, MD: National Asphalt Pavement Association, No. BMD 508.

Hayes, A.F. and Cai, L., 2007. Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, 39 (4), 709–722.

Im, S. and Zhou, F., 2017. New and Simpler Cracking Test Method for Asphalt Mix Designs. *Transportation Research Record: Journal of the Transportation Research Board*, 2631 (1), 1–10.

Jarque, C.M. and Bera, A.K., 1980. Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Economics Letters*, 6 (3), 255–259.

Kumar, V., Coleri, E., and Sukhija, M., 2024. Assessing the Suitability of Different Testing and Specimen Preparation Approaches for Indirect Tensile Cracking Test (IDT-CT) Using a Simple Ranking Framework. *Journal of Testing and Evaluation*, 52 (5), 2761–2773.

Leavitt, A., Treybig, C., Epps-Martin, A., and Arambula-Mercado, E., 2024. Correlation Between Laboratory Testing and Field Performance for Balanced Mix Design of Asphalt Pavements. *In*: A. Carter, K. Vasconcelos, and E. Dave, eds. *14th International Conference on Asphalt Pavements ISAP2024*. Presented at the International Society for Asphalt Pavements, Montreal, CA: Springer Nature Switzerland, 67–71.

Lemke, Z., Sadek, H., Swiertz, D., Reichelt, S., and Bahia, H.U., 2018. Effects of Reheating Procedure and Oven Type on Performance Testing Results of Asphalt Mixtures. *Transportation Research Record*, 2672 (26), 124–133.

Long, J.S. and Ervin, L.H., 2000. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 54 (3), 217–224.

Montgomery, D.C., 2019. *Design and Analysis of Experiments*. 10th ed. Hoboken, NJ: John Wiley & Sons, Inc.

Montgomery, D.C. and Runger, G.C., 2018. *Applied Statistics and Probability for Engineers*. 7th ed. Hoboken, NJ: John Wiley & Sons, Inc.

Mousavi Rad, S., Olagunju, M., and Elkashef, M., 2025. Impact of Binder and Aggregate Type on the Cracking, Rutting, and Moisture Susceptibility of Asphalt Mixes in Oklahoma: A BMD Approach. *Journal of Materials in Civil Engineering*, 37 (3), 04024530.

Newcomb, D., Epps Martin, A., Yin, F., Arambula, E., Park, E.S., Chowdhury, A., Brown, R., Rodezno, C., Tran, N., Coleri, E., Jones, D., Harvey, J.T., and Signore, J.M., 2015. *NCHRP Report 815: Short-Term Laboratory Conditioning of Asphalt Mixtures*. Washington D.C.: National Cooperative Highway Research Program, No. NCHRP Report 815.

Pasandín, A.R. and Pérez, I., 2014. Effect of ageing time on properties of hot-mix asphalt containing recycled concrete aggregates. *Construction and Building Materials*, 52, 284–293.

Pei, Y., Bahia, H.U., Zhang, R., Wang, R., and Xie, R., 2025. Revisiting Asphalt Mixture Cracking Performance: Development of a Work of Fracture-Based Crack Propagation Resistance Index from IDEAL-CT. *Engineering Fracture Mechanics*, 328, 111513.

Rajan, B., Suchismita, A., and Singh, D., 2023. Rutting Resistance Evaluation of Highly Polymer-Modified Asphalt Binder and Mixes Using Different Performance Parameters. *Journal of Materials in Civil Engineering*, 35 (8), 04023266.

Shingala, M.C. and Rajyaguru, A., 2015. Comparison of Post Hoc Tests for Unequal Variance. *International Journal of New Technologies in Science and Engineering*, 2 (5), 22–33.

Smith, B., 2017. Characterizing Short and Long Term Mixture Aging of a Full-Scale and Non-Trafficked Asphalt Test Section. *Theses and Dissertations*.

Taylor, A., 2025. Effect of Dwell Time and Lag Time on IDEAL-CT and IDEAL-RT Results - Expanded Study.

Tiwari, A., 2026. Comparative Study on the Long-Term Performance of BMD Mixtures Prepared Using Different Sample Preparation Techniques and Superpave Mixtures Under Varying Pavement Structures - ProQuest [online]. Available from: https://www.proquest.com/openview/286598390839eecf7b95ca36d7130d27/1?cbl=18750&diss=y&pq-origsite=gscholar [Accessed 8 Jun 2026].

Vamsikrishna, G. and Singh, D., 2023. Comparison of Rutting Resistance of Plant Produced Asphalt Mixes Using Hamburg Wheel Tracker and Surrogate Simple Performance Tests: IDEAL-RT and HT-IDT. *Journal of Materials in Civil Engineering*, 35 (12), 04023471.

Veginati, V., Mensching, D.J., Elwardany, M.D., Andriescu, A., Haghshenas, H., and Jahangiri, B., 2025. *Evaluating the Sensitivity of Intermediate Temperature Performance Tests With Short-Term and Long-Term Laboratory Conditioning: Validation Study*. Washington D.C.: Federal Highway Administration, No. FHWA-HRT-24-147.

Zhou, F., Crockford, B., Zhang, J., Hu, S., Epps, J., and Sun, L., 2019. Development of an IDEAL Rutting Test for Asphalt Mixture Design, Quality Control and Quality Assurance. *Association of Asphalt Paving Technologists*, 88, 719–750.

Zhou, F., Hu, S., and Newcomb, D., 2020. Development of a Performance-Related Framework for Production Quality Control with Ideal Cracking and Rutting Tests. *Construction and Building Materials*, 261, 120549.

Zhou, F., Im, S., Sun, L., and Scullion, T., 2017. Development of an IDEAL cracking test for asphalt mix design and QC/QA. *Road Materials and Pavement Design*, 18 (sup4), 405–427.

Zhou, F., Mogawer, W.S., Elwardany, M., Mensching, D.J., and Mahmoud, E., 2025. Development of a Laboratory Operations-Friendly Long-Term Loose Mix Aging Protocol for New Asphalt Pavements. *Transportation Research Record*, 2679 (1), 679–694.

Data Availability Statement

The summary data and analysis for this study are hosted at https://github.com/TFHRC-ABML/BMD_Specimen_Fabrication. Raw data files are available upon request from Aaron Leavitt (aaron.leavitt@dot.gov).

Appendix 1. Statistical Analysis Methodology

This appendix details the statistical framework utilized to assess the influence of specimen fabrication methods on BMD performance testing outcomes. To ensure robust and reliable results, the IDEAL-CT, IDEAL-RT, and HWTT datasets were first subjected to systematic outlier detection algorithms to identify and filter anomalous data points. Subsequently, comparative evaluations were conducted using a suite of statistical tools, including Student's t-tests, one-way and two-way Analyses of Variance (ANOVA), and Tukey's Honestly Significant Differences (HSD) post-hoc analysis. The specific theoretical methodologies and outlier detection protocols governing these analyses are outlined below.

A1.1 Outlier Detection

Identifying and treating statistical outliers is a critical preliminary step before performing the statistical analysis on the collected database. Extreme values, whether stemming from material anomalies or testing irregularities, can disproportionately influence parametric statistical models and obscure the true effects of specimen fabrication methods. To ensure the integrity of the database while maintaining statistically viable sample sizes, distinct outlier detection protocols were employed for the tests in first three tasks (IDEAL-CT and IDEAL-RT) and those in the last task (HWTT). These protocols were specifically tailored to the unique data structures and subsequent analytical requirements of each test.

A1.1.1 IDEAL-CT and IDEAL-RT Tests

The dataset for the IDEAL-CT and IDEAL-RT test results initially consisted of 7 and 5 replicates per sample set, respectively. Outlier detection for these tests was performed directly on the calculated performance metrics (CT-Index and RT-Index) using a rigorous, two-step methodology:

Step 1: Intra-set Z-Score Screening

An initial screening was conducted within each sample set to identify extreme deviations from the mean. For each replicate, a standard *z*-score was calculated (Montgomery and Runger 2018) using Equation $A1.1$.

$$z = \frac{x_i - \mu}{\sigma} \qquad (A1.1)$$

where x_i is the individual CT- or RT-index value, μ is the average of the index value for given sample set, and σ is the standard deviation of the index value for the given sample set. Any datapoint yielding an absolute *z*-score of two or greater ($|z| \geq 2$) was flagged as an outlier and excluded from further statistical analysis. However, to ensure statistical reliability for the downstream analysis, a strict boundary condition was enforced: a minimum of three replicates per sample set was always maintained.

To quantify the impact of the inter-set z-score screening on the dataset, Table A1.1 summarizes the data retention metrics for both the IDEAL-CT and IDEAL-RT tests across the evaluated tasks. As this table shows, while 29 anomalous data points were successfully identified and removed (out of 1,134), the strict boundary condition was maintained, ensuring no sample set fell below the required minimum of three replicates.

Table A1.1. Summary of dataset replicates before and after *z*-score screening.

Task	Test	Number of datapoints		Number of datapoints per sample set		
		Including Outliers	Excluding Outliers	Minimum	Average	Maximum
Task 1	IDEAL-CT	252	244	6	6.78	7
	IDEAL-RT	90	90	5	5.00	5
Task 2	IDEAL-CT	182	175	6	6.73	7
	IDEAL-RT	130	130	5	5.00	5
Task 3	IDEAL-CT	280	267	6	6.67	7
	IDEAL-RT	200	200	5	5.00	5
All	Both	1,134	1,105	5	5.91	7

Step 2: Residual Normality Verification

Because the core comparative analysis relied on two-way ANOVA, the dataset was required to meet the parametric assumption of normal distribution. A Jarque-Bera (JB) goodness-of-fit test was applied to the residuals generated by the two-way ANOVA model (Jarque and Bera 1980). In instances where the residuals failed the normality check, an iterative trimming process was initiated. The data point corresponding to the highest absolute residual was removed, and the ANOVA model and JB test were recalculated. This iterative removal continued solely until the residual distribution passed the JB normality test. As in Step 1, this process was strictly constrained to ensure that no sample set was reduced to below three replicates.

Following the iterative residual trimming process, the assumption of normality for the two-way ANOVA models was successfully satisfied. Table A1.2 summarizes the final dataset dimension and the quantitative results of the JB normality check evaluations. The table presents the final JB test statistics and corresponding p-values (where p-value > 0.05 indicates a normal distribution of residuals).

Table A1.2. Summary of JB normality test results and final dataset dimensions.

Task	Test	Number of Sample Sets	JB Normality Statistic		Number of Sample Sets with X Replicates				
			p-value	F-stat	3	4	5	6	7
Task 1	IDEAL-CT	36	0.33	2.21	-	-	-	8	28
	IDEAL-RT	18	0.63	0.91	-	2	16	-	-
Task 2	IDEAL-CT	26	0.23	2.93	-	1	2	8	15
	IDEAL-RT	26	0.09	4.76	1	2	23	-	-
Task 3	IDEAL-CT	40	0.06	5.53	-	1	4	16	19
	IDEAL-RT	40	0.24	2.89	-	-	40	-	-

A1.1.2 HWTT Test

The HWTT dataset presented a different analytical structure, providing continuous rutting curves rather than single discrete indices, with an initial count of four replicates per sample set. To apply standard outlier detection metrics to this continuous rutting data, the curves were first quantified by calculating the Area Under the Curve (AUC) for each of the four replicates. Following this quantification, the standard *z*-score screening protocol (Montgomery and Runger 2018) was applied directly to the calculated AUC values, retaining only those where the absolute *z*-score was less than two ($|z| \leq 2$), to identify anomalous rutting behavior within each sample set. Utilizing this approach across the 21 unique sample sets (7 mixtures × 3 trimming methods), no outliers were detected in 6 of the sample sets. In each of the remaining 15 sample sets, exactly one of the four replicates was identified as an outlier.

However, in standard practice, highway agencies rarely remove outliers from HWTT data. Furthermore, the baseline replicate count of four inherently limits the robustness of aggressive data filtering. To account for these practical constraints while still rigorously assessing the data, a dual-analysis approach was adopted, where all statistical evaluations for the HWTT were conducted twice: once utilizing the raw, unfiltered dataset, and a second time utilizing the filtered dataset with the AUC-based outliers removed.

A1.2 Student's t-Test

The Student's two-sample t-test was utilized as the primary inferential statistical tool to evaluate differences in mean CT- or RT-indices between two distinct categorical groups. This parametric test assesses whether the observed differences between sample means are statistically significant or if they are likely attributable to random testing variation. The test calculates a t-statistic by quantifying the difference between the group means relative to the dispersion (variance) within the groups, as calculated (Montgomery and Runger 2018) using Equation $A1.2$:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{A1.2}$$

where μ_i represents the average of the i^{th} sample set, σ_i represents the standard deviation of the i^{th} sample set, and n_i represents the number of datapoints within i^{th} sample set, respectively. The Student's t-test was specifically implemented in Tasks 1 and 2 to investigate the isolated effects of distinct specimen fabrication variables on mixture performance. Because the experimental design for these specific tasks evaluated exactly two categorical fabrication states per variable (G_{mm} aging level in Task 1 and specimen conditioning methods in Task 2), the Student's two-sample t-test served as the optimal statistical tool for comparative analysis. For each mixture at given aging level, the calculated CT- or RT-indices under the two respective conditions were directly compared to yield a t-statistic and a corresponding p-value. A standard confidence threshold ($\alpha = 0.05$) was then applied to the p-value to determine the statistical significance of the observed differences, assuming a null hypothesis of no true difference between the two fabrication methods being compared.

To interpret the comparative results across different mixtures and aging levels, a multidimensional graphical approach was adopted by plotting the calculated t-statistics directly against their corresponding p-values. This visualization effectively synthesizes the data by using a p-value threshold line (at p-value 0.05) to segregate statistically significant treatment effects from random testing noise. Concurrently, the t-statistic axis provides critical insight into both the magnitude and directionality of the observed effect, as its positive or negative sign reveals whether a specific conditioning protocol systematically increased or decreased the performance index relative to the baseline.

A1.3 One-Way ANOVA

For Tasks 3 and 4, the experimental design expanded beyond binary comparisons to encompass more than two categorical specimen fabrication methods (four heating methods in task 3 and three cutting methods in task 4). Since performing multiple, unadjusted t-tests across several groups increases the probability of Type I errors (false positives), a one-way ANOVA was implemented as the appropriate primary analytical tool (Montgomery 2019). The one-way ANOVA is a parametric statistical method used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. It accomplishes this by comparing the CT- or RT-index variation observed between the different fabrication methods to the variation observed within each individual method's replicate sets. This comparison yields an F-statistic, conceptually represented in Equation $A1.3$:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{A1.3}$$

where $MS_{between}$ represents the mean square variance between the group means, and MS_{within} represents the mean square error (the variance within the groups due to random testing noise). By evaluating the calculated F-statistic against the F-distribution, a corresponding p-value is generated. Under a standard confidence threshold ($\alpha = 0.05$), a significant p-value indicates that the null hypothesis (which states that all fabrication methods produce identical average CT- or RT-indices) must be rejected. Therefore, a significant ANOVA result confirms that at least one of the specimen fabrication methods exerted a fundamental, statistically significant effect on the mixture's measured performance. However, because the one-way ANOVA is an omnibus test, it solely identifies the presence of an overall effect; while it does not specify which particular fabrication methods significantly differ from one another. Consequently, achieving a significant p-value in this step serves as the prerequisite gateway for conducting subsequent post-hoc evaluations (Montgomery 2019).

A1.4 Post-Hoc Analysis

Following a statistically significant omnibus result from an ANOVA, it is typically standard practice to employ Tukey's Honestly Significant Difference (HSD) post-hoc test to determine precisely which specimen fabrication methods differ from one another. However, Tukey's HSD strictly relies on the assumption of homogeneity of variance across all groups. Since such assumption is not valid for some portion of dataset in this study, as it will be explained later in Section A1.6, the variance among the testing groups was significantly heterogeneous (heteroscedasticity), applying Tukey's method would risk compounding the Type I error rate (Shingala and Rajyaguru 2015). Consequently, the Games-Howell post-hoc test was selected as the appropriate analytical approach.

The Games-Howell method is a non-parametric post-hoc analysis designed specifically for multiple comparison scenarios where the assumptions of equal variances and equal sample sizes are violated. Like Tukey's HSD, it conducts exhaustive pairwise comparisons between all possible combinations of group means while strictly controlling the overall family-wise error rate. However, instead of relying on a single, pooled intra-group variance (mean square error) for the entire dataset, the Games-Howell test dynamically calculates a unique standard error and distinct degrees of freedom for every single pair being compared.

The mathematical formulation for the critical difference threshold between any two specific fabrication methods, *i* and *j*, is generally expressed as shown in Equation $A1.4$ (Games and Howell 1976):

$$GH_{ij} = q_{(k, df_{ij}, \alpha)} \sqrt{\frac{1}{2}\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)} \tag{A1.4}$$

where *q* represents the critical value derived from the Studentized range distribution (a normalized probability distribution used specifically to control false positive rates during multiple comparisons) for *k* groups at the selected α level, s_i^2 and s_j^2 are the individual variances of the two respective groups, and n_i

and n_j are their respective replicate counts. The degrees of freedom (df_{ij}) for each pairwise comparison are calculated independently using the Welch-Satterthwaite equation.

During the evaluation of Tasks 3 and 4, the absolute difference between the average performance metrics of any two specific fabrication methods was directly compared against this dynamically calculated GH_{ij} threshold. If the absolute difference between the group means exceeded the Games-Howell value, those two fabrication methods were deemed to have a statistically significant, distinct impact on mixture performance. Conversely, differences smaller than the threshold indicated statistical parity between those specific methods, accounting for their unequal variances.

A1.5 Two-Way ANOVA

While the one-way ANOVA and Student's t-tests were utilized to evaluate isolated variables, the two-way Analysis of Variance (ANOVA) served as the primary, comprehensive analytical model across all experimental tasks. The two-way ANOVA is a robust parametric methodology designed to evaluate the influence of two independent categorical variables concurrently on a single continuous dependent variable (Montgomery 2019). In the context of this study, the two independent variables established for the model were the specific asphalt mixture designation (Mix ID) and the applied specimen fabrication method, with the calculated performance indices (e.g., CT-index, RT-index, flexibility, G_f, CPR, etc.) serving as the dependent response variables.

The implementation of the two-way ANOVA provides two distinct tiers of statistical insight. First, it evaluates the "main effects" of the independent variables. It systematically determines whether the sample variance is significantly driven by the fundamental differences between the asphalt mixtures themselves, and, crucially for the objective of this study, whether the fabrication method introduces a statistically significant variation in the overall testing outcomes regardless of the mixture type.

Beyond the main effects, the defining advantage of the two-way ANOVA is its capacity to quantify the "interaction effect" between the two independent variables. By analyzing the interaction term (Mix ID × Fabrication Method), the model statistically assesses whether the influence of a given fabrication protocol depends on the specific mixture being tested. A statistically significant interaction indicates a non-uniform response, revealing that certain mixture designs may be highly sensitive to deviations in specimen fabrication procedures, while other mixtures remain relatively stable under the same variations. Consequently, achieving an understanding of specimen fabrication effects requires interpreting both the main effect of the fabrication methods and their specific interaction with the volumetric and material properties of each asphalt mixture.

A1.6 ANOVA Assumptions and Robust Standard Errors (HC3)

The validity of both one-way and two-way ANOVA models relies heavily on two fundamental parametric assumptions: the normal distribution of model residuals and the homogeneity of variances (homoscedasticity) across all compared groups (Montgomery 2019). As detailed in the earlier discussion on outlier detection protocols (section A1.1 Outlier Detection), the assumption of normality was inherently addressed and satisfied for Tasks 1 through 3 via the iterative residual trimming process. Consequently, the database utilized for the inferential analyses successfully met the normality prerequisite. To evaluate the second critical assumption, Levene's test was conducted to assess the equality of variances across the different sample sets. Levene's test is a robust inferential statistic used to evaluate the null hypothesis that population variances are equal (Montgomery 2019), and it is widely preferred in engineering research because it is less sensitive to minor departures from perfect normality than alternative variance tests.

A comprehensive summary of the Levene's test statistics and their corresponding p-values for each task and test type is provided in Table A1.3. As this table shows, the outcomes for the CT-indices from IDEAL-CT test results yielded statistically significant p-values ($p \leq 0.05$) across the evaluated datasets, which indicates the rejection of the null hypothesis. This outcome confirmed that the assumption of homogeneity of variance was not satisfied and revealed significant heteroscedasticity within the database.

Table A1.3. Summary of Levene's Homogeneity Test Results.

Task	**Test**	**Number of Sample Sets**	**Levene's test**	
			p-value	**F-stat**
Task 1	IDEAL-CT	36	0.00	7.98
	IDEAL-RT	18	0.09	0.54
Task 2	IDEAL-CT	26	0.01	1.85
	IDEAL-RT	26	0.33	1.12
Task 3	IDEAL-CT	40	0.00	2.22
	IDEAL-RT	40	0.39	1.06

When faced with a violation of homoscedasticity, particularly within an unbalanced experimental design stemming from outlier removal (where number of replicates in comparing sample sets can be different), standard ANOVA models are at risk of calculating skewed F-statistics and generating inaccurate p-values. To rectify this, two primary corrective methodologies are typically considered: applying a mathematical data transformation or utilizing robust standard errors. Data transformation, such as applying a natural logarithm to the performance values, introduces complexity to the data and inherently obscures the physical interpretation of the engineering results.

Therefore, to resolve the homogeneity violation while strictly preserving the original engineering units of measurement, the Heteroskedasticity-Consistent Standard Errors Type 3 (HC3) method was implemented. Rather than altering the underlying data points, the HC3 estimator corrects the mathematical foundation of the ANOVA by dynamically adjusting the standard errors and applying weight factors according to the specific variance observed within each individual group. This estimation technique is specifically designed to control the Type I error rate and generate accurate p-values in datasets characterized by small, unbalanced sample sizes and unequal variances. By integrating the HC3 calculation method directly into both the one-way and two-way ANOVA models, the statistical framework effectively neutralized the heteroscedasticity issue, ensuring that the final comparative assessments of the specimen fabrication methods remained mathematically rigorous and practically interpretable (Long and Ervin 2000, Hayes and Cai 2007).

Appendix 2: Analytical Background

This appendix details the comprehensive analytical background and methodologies employed to process the raw load-displacement data collected during the IDEAL-RT and IDEAL-CT performance testing. The following sections document the specific testing standards, theoretical principles, and mathematical formulations utilized to calculate the suite of performance metrics evaluated in this study. These metrics include RT-index and tensile strength for IDEAL-RT test, and CT-index, flexibility parameter, failure energy (G_f) and cracking propagation resistance (CPR) for IDEAL-CT test.

A2.1 Analysis of IDEAL-RT Test Results

For the IDEAL-RT test, the raw data generated by the data acquisition systems typically consists of three primary continuous arrays: time, applied load, and actuator displacement. Additionally, the geometric properties including diameter and thickness are recorded prior to testing.

Initially, the recorded time and actuator displacement data are used as a quality control measure to verify the applied displacement rate. This check ensures that the actual loading rate remains within the acceptable tolerance range of 50.0 ± 2.0 mm/min throughout the duration of the test, as required by the ASTM D8360 standard (ASTM International 2022).

Following this verification, the primary performance metric, the RT-index, is calculated utilizing the standardized methodology specified in ASTM D8360 (ASTM International 2022). First, the shear strength (τ_f) of the mixture is calculated from the recorded peak load using Equation $A2.1$.

$$\tau_f = 0.356 \times \frac{P_{max}}{t \times w} \tag{A2.1}$$

where τ_f is the shear strength (Pa), P_{max} is the maximum applied load (N), t is the specimen thickness (m), and w is the width of the upper loading strip (established as 0.0191 m for standard 150-mm specimen). Subsequently, the RT-index is derived from the calculated shear strength using Equation $A2.2$.

$$RT_{Index} = 6.618 \times 10^{-5} \times \frac{\tau_f}{1\ Pa} \tag{A2.2}$$

where $1\ Pa$ serves as a unit cancellation factor and 6.618×10^{-5} is the standardized scaling coefficient.

Finally, in addition to the standard RT-index, the indirect tensile strength is calculated as a supplemental output parameter. The calculation of this metric is inspired by the indirect tensile test (IDT) methodology outlined in AASHTO T 322-07 (AASHTO 2011), which defines the tensile strength of a cylindrical specimen using Equation $A2.3$.

$$\sigma = \frac{2 \times P_{max}}{\pi \times t \times D} \tag{A2.3}$$

where σ is the tensile strength (Pa), D is the specimen diameter (m), and other parameters were previously defined.

A2.2 Analysis of IDEAL-CT Test Results

For the IDEAL-CT test, the data acquisition system typically generates three primary arrays of continuous raw data: time, applied load, and load-line displacement. Alongside these arrays, thickness and diameter are recorded prior to testing.

As an initial quality control measure, the recorded time and displacement arrays are analyzed to verify the actual applied loading rate. This step is critical to ensure that the test was executed accurately and that the displacement rate remained within the strict tolerance of 50.0 ± 2.0 mm/min, in accordance with ASTM D8225 specifications (ASTM International 2019).

Once the loading rate is verified, a suite of performance metrics is calculated. First, the standardized parameters defined in ASTM D8225 are extracted. The work of failure (W_f) is calculated as the total area under the load-displacement curve using a trapezoidal numerical integration. From this, the failure energy (G_f) is determined by dividing the work of failure by the cross-sectional area of the specimen using Equation $A2.4$.

$$G_f = \frac{W_f}{D \times t} \tag{A2.4}$$

where G_f is the failure energy (J/m^2), W_f is the work of failure (J), D is the specimen diameter (m), and t is the specimen thickness (m). Figure A2.1 illustrates these parameters for one of the IDEAL-CT test specimens (T1_L5_M1_S11.csv).

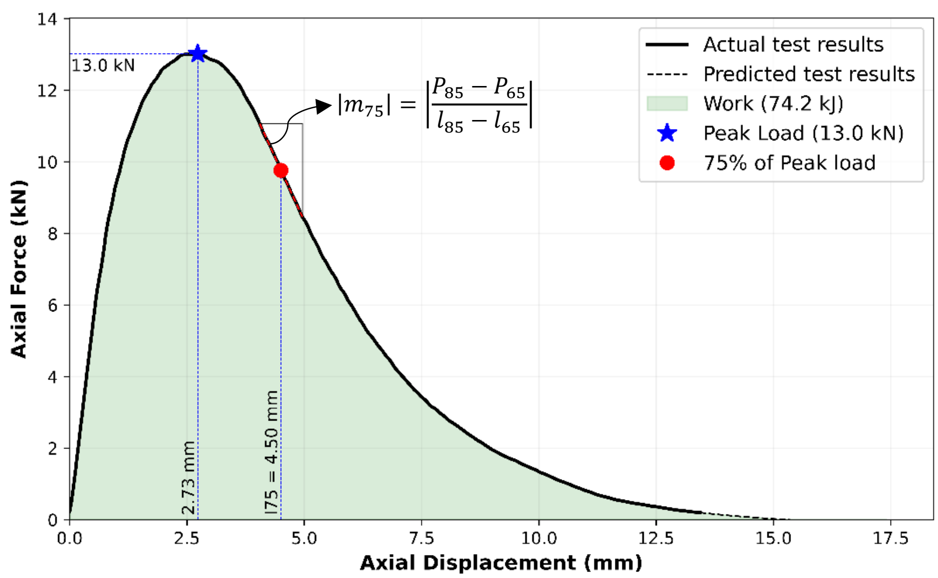


Figure A2.1. Example illustration of the IDEAL-CT raw data analysis for T1_L5_M1_S11 specimen (source: FHWA).

Next, the flexibility parameter is calculated as the ratio of the post-peak displacement to the absolute slope at the 75% peak load point using Equation $A2.5$. These values are then compiled to calculate the primary cracking metric, the CT-index using Equation $A2.6$.

$$Flexibility = \frac{l_{75}}{|m_{75}|} \tag{A2.5}$$

$$CT_{Index} = \frac{t}{0.062\ m} \times \frac{l_{75}}{D} \times \frac{G_f}{|m_{75}|} \qquad (A2.6)$$

Where l_{75} is the post-peak displacement at the 75% peak load point (m), $|m_{75}|$ is the post-peak absolute slope at the 75% peak load point (N/m), $0.062\ m$ serves as a thickness normalization constant, and all other parameters were previously defined.

In addition to the standard ASTM D8225 metrics, the indirect tensile strength (σ) of the mixture is calculated as a fundamental physical property. Identical to the methodology applied for the IDEAL-RT, this calculation is inspired by AASHTO T 322 using the peak applied load and the specimen dimensions, as shown in Equation $A2.3$.

Finally, the Cracking Propagation Resistance (CPR) parameter is calculated as it was developed claiming to address the limitations of the single-point slope dependency in the CT-index parameter (Pei *et al.* 2025). The CPR metric evaluates the average cumulative energy required for crack propagation by analyzing the entire post-peak stage. For this purpose, the raw data was converted into the cumulative work of failure as a function of actuator displacement. As an example, Figure A2.2 shows this curve for the T1_L5_M1_S11 specimen. The CPR parameter integrates the area under this curve strictly during the crack propagation phase—defined from the slope inflection point (the peak load) to the test endpoint (where the load drops to 0.1 kN). Finally, the CPR is calculated using Equation $A2.7$.

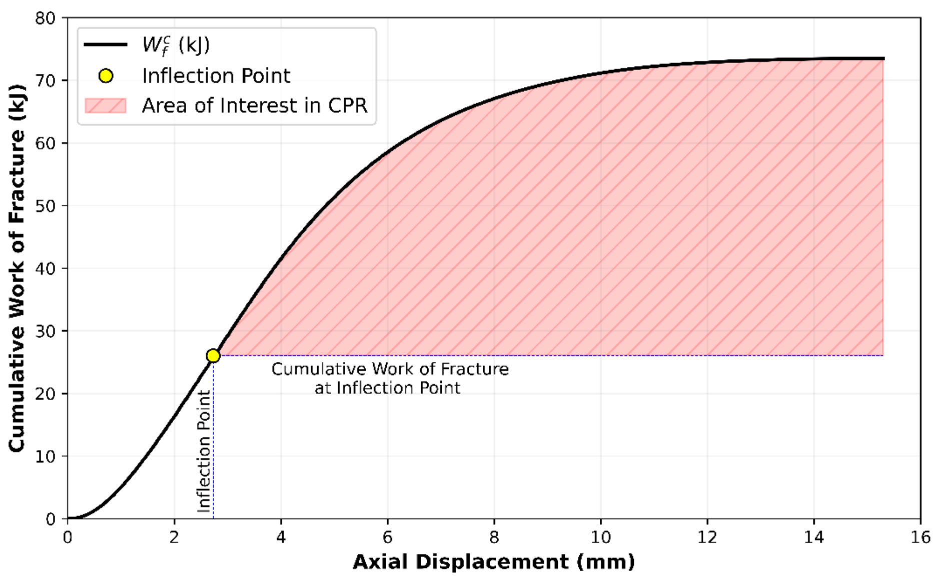


Figure A2.2. Example illustration of the CPR calculation from IDEAL-CT cumulative work of failure as a function of displacement curve for T1_L5_M1_S11 specimen (source: FHWA).

$$CPR = \frac{t}{0.062\ m} \times \frac{D}{0.150\ m} \times \frac{1}{\Delta Disp} \times \left[\int_{l_{peak}}^{l_{end}} W_f^c(l) \cdot dl - \Delta Disp \times W_{f,inflection}^c \right] \qquad (A2.7)$$

$$\Delta Disp = l_{end} - l_{peak} \qquad (A2.8)$$

Where CPR is cracking propagation resistance parameter, $W_f^c(l)$ is the cumulative work of failure at certain displacement (J), $W_{f,inflection}^c$ is the cumulative work of failure corresponding to the point of peak load (J), l_{peak} is the displacement corresponding to the peak load (m), and l_{end} is the displacement at the endpoint, where the load reduces to 0.1 kN (m).

A2.3 Analysis of HWTT Test Results

For the HWTT, the raw data collected during testing typically consists of two primary continuous arrays: the number of applied wheel passes and the corresponding measured rut depth. To process and efficiently interpret these raw rutting curves, this study utilized AutoHWTT software, an open-source analytical software tool developed at the Federal Highway Administration's (FHWA's) Asphalt Binder and Mixtures Laboratory (ABML) specifically to automate HWTT data analysis and ensure user-independent results. The software's analysis engine employs the newly proposed Two-Part Power (2PP) model, which provides a unified, continuous, and differentiable mathematical representation of the entire HWTT rutting curve. The 2PP model combines two distinct power-law functions separated by a specific transition boundary. This model can be mathematically expressed as shown in Equation $A2.9$.

$$TRD(N) = \begin{cases} a \times N^b & \text{for } N \leq SN \\ \alpha \times (N - \gamma)^{\beta} + \varphi & \text{for } N > SN \end{cases} \qquad (A2.9)$$

where *TRD(N)* is the total rut depth at load pass *N*; *a*, *b*, *α*, *β*, *γ*, and *φ* are calibration coefficients; and *SN* is the stripping number. The model applies strict continuity and consistency boundary conditions at the transition point ($N = SN$), which effectively reduces the required fitting coefficients to four and ensure smooth, robust mathematical transition between the viscoplastic and stripping-induced damage phases (Abdollahi *et al.* TBD). Once the 2PP model is fitted to the raw data, a comprehensive suite of rutting and moisture susceptibility parameters is extracted, as explained below. In addition, Figure XX

- **Total Rut Depth (TRD)**: the absolute maximum rut depth recorded at test completion, representing the combined deformation from both mechanical loading (viscoplasticity) and potential stripping.
- **Corrected Rut Depth at 20,000 passes (CRD20)**: a parameter designed to isolate the viscoplastic deformation resistance of the mixture by entirely excluding stripping-induced deformation. It is calculated using only the primary segment of the 2PP model ($CRD = a \times N^b$) evaluated at 20,000 passes.
- **Stripping Number (SN)**: the specific number of load passes corresponding to the mathematical inflection point of the fitted 2PP curve. This signifies the boundary point between the secondary creep phase and the onset of the tertiary stripping phase.
- **Stripping Inflection Point (SIP)**: the conventional moisture susceptibility metric, determined as the intersection of the projected creep slope line and the stripping slope line drawn from the test's end point.
- **Adjusted Stripping Inflection Point (Adjusted SIP)**: a refined SIP calculation where the stripping slope is evaluated based on an end point fixed at a 12.5 mm rut depth threshold, rather than the arbitrary end of the test. This adjustment incorporates the magnitude of rutting into the SIP estimation, providing a more reliable assessment and preventing the rejection of mixtures that exhibit a stripping phase but maintain low overall deformation.

To visualize this analytical approach, Figure A2.3 provides a representative example of the HWTT data. As depicted in the figure, the 2PP model is fitted directly to the raw experimental data, allowing for the clear geometric and mathematical extraction of all key performance parameters, including the SN, conventional SIP, Adjusted SIP, and CRD20.

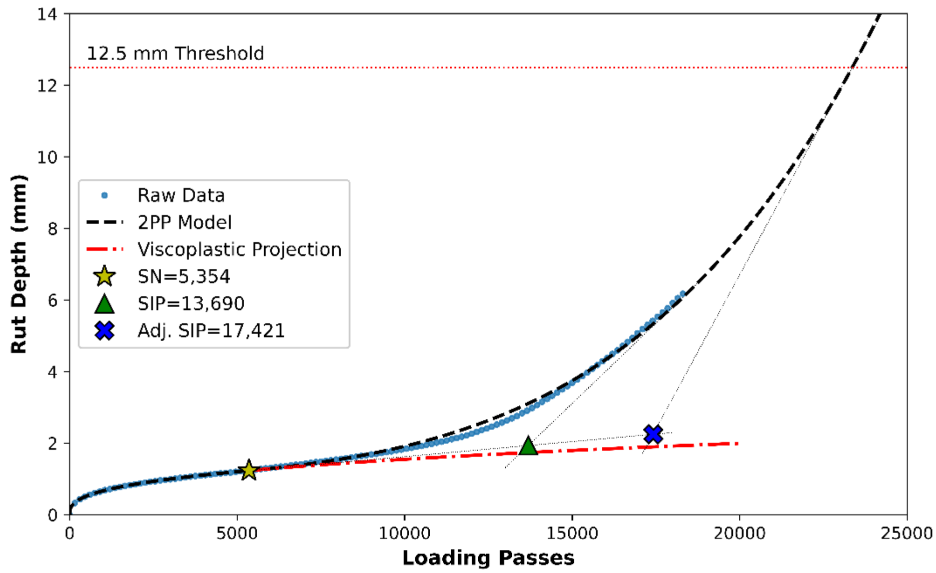


Figure A2.3. Illustration of the 2PP model fitted to HWTT raw highlighting the extraction of key viscoplastic and stripping parameters (source: FHWA).

Appendix 3: Comprehensive Analytical Results for Task 1

This appendix provides the supplementary volumetric data, figures, and comprehensive statistical outputs supporting the Task 1, which focuses on evaluating the effect of G_{mm} aging level on IDEAL-CT and IDEAL-RT performance test results.

A3.1 Effect of G_{mm} Aging Level on Air Voids

As detailed earlier in the body of the report (see Figure 3), LTOA aging of the nine evaluated mixtures resulted in an increase in the G_{mm} across all sample sets, with a maximum observed increase of 4.68%. Notably, when excluding the highly sensitive L2-M2 mixture, this densification was strictly limited to an increase of 0.52%. To further understand the practical implications of these density shifts, the effect of the G_{mm} aging level on the volumetric properties of the compacted specimens was systematically evaluated.

For this purpose, test specimens that were originally compacted targeting specific densities based on their LTOA G_{mm} values were selected, and their measured bulk specific gravity (G_{mb}) values were retrieved. To quantify the volumetric discrepancy, the air voids for these physical specimens were then recalculated utilizing their measured G_{mb} in conjunction with the corresponding STOA G_{mm} values. This recalculation was performed using the standard volumetric relationship as shown in Equation $A3.1$, and the results are shown in Figure A3.1.

$$V_a = \left(1 - \frac{G_{mb}}{G_{mm}}\right) \times 100 \qquad (A3.1)$$

where V_a is the air voids (%), G_{mm} is the theoretical maximum specific gravity and G_{mb} is the bulk specific gravity.

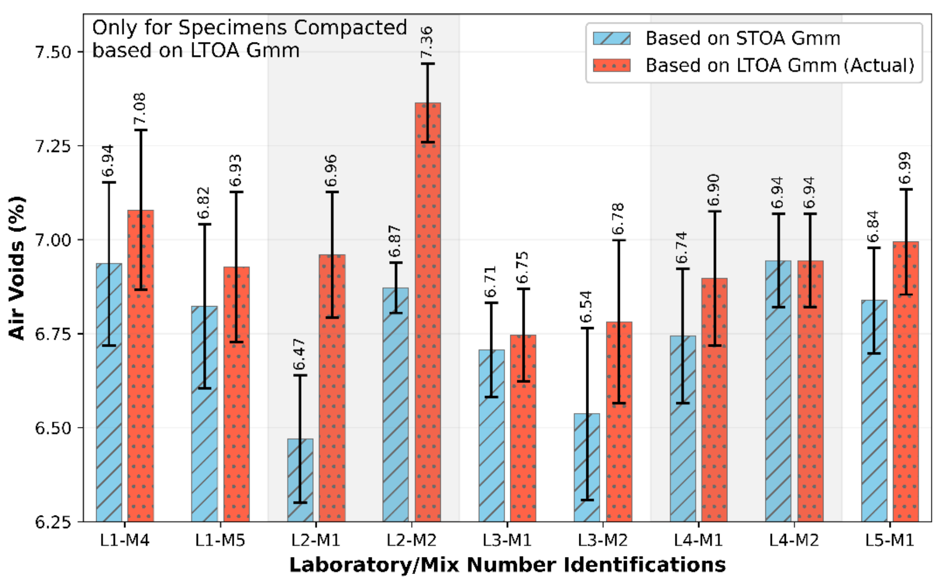


Figure A3.1. Actual and recalculated air voids of the specimens compacted based on the LTOA G_{mm} values (source: FHWA).

Figure A3.1 illustrates a comparative analysis of the recalculated air voids for the test specimens under both G_{mm} baseline conditions. In this figure, the red bars denote the actual, measured air voids of the specimens that were compacted targeting specific densities using their respective LTOA G_{mm} values as the reference maximum density. Conversely, the blue bars represent the recalculated air voids for those identical physical specimens if their unaged STOA G_{mm} values were retroactively applied for air void calculations. A comparative assessment of these two datasets reveals that the aging-induced shifts in G_{mm} have a relatively limited impact on the final air void determinations. Even for mixture L2-M2, which previously exhibited the highest G_{mm} percentage increase, the corresponding translation to volumetric variance remains minimal. Specifically, the data demonstrates that the choice between STOA and LTOA G_{mm} baselines alters the calculated specimen air voids by a maximum of only 0.5% across all investigated mixtures.

To further evaluate the broader distribution and consistency of the specimen air voids, Figure A3.2 shows an equality plot comparing the mean measured air voids of the sample sets compacted using the STOA G_{mm} against those compacted using the LTOA G_{mm}. When interpreting this specific visualization, it is critical to note that the physical specimens comprising the STOA G_{mm} dataset are distinct from those comprising the LTOA G_{mm} dataset. Because these represent independent specimen fabrication efforts rather than a retroactive recalculation on a single set of samples, the graph does not function as a strict 1:1 paired equality assessment. Rather, the primary utility of this plot is to verify the accuracy of the fabrication process across both density baselines. As this figure shows, despite the different reference G_{mm} values and the inherent variability of laboratory compaction, the mean air voids and their associated replicate error bars for all sample sets consistently fall within the required target tolerance of 7.0% ± 0.5%.

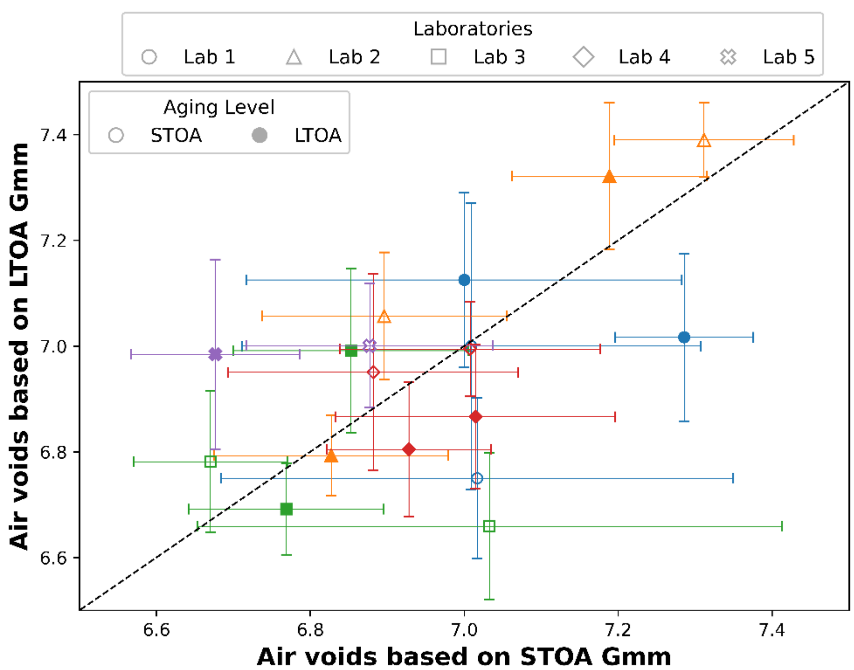


Figure A3.2. Air voids distribution equality graphs for test specimens used in Task 1 (source: FHWA).

A3.2 Equality Plots on the Effect of G_{mm} Aging Level

This section includes the equality plots showing the effect of G_{mm} aging level on different parameters calculated for IDEAL-CT and IDEAL-RT tests. Each datapoint on these plots indicates the average parameter value for the sample set compacted targeting STOA G_{mm} values (on horizontal axis) and LTOA G_{mm} values (on vertical axis).

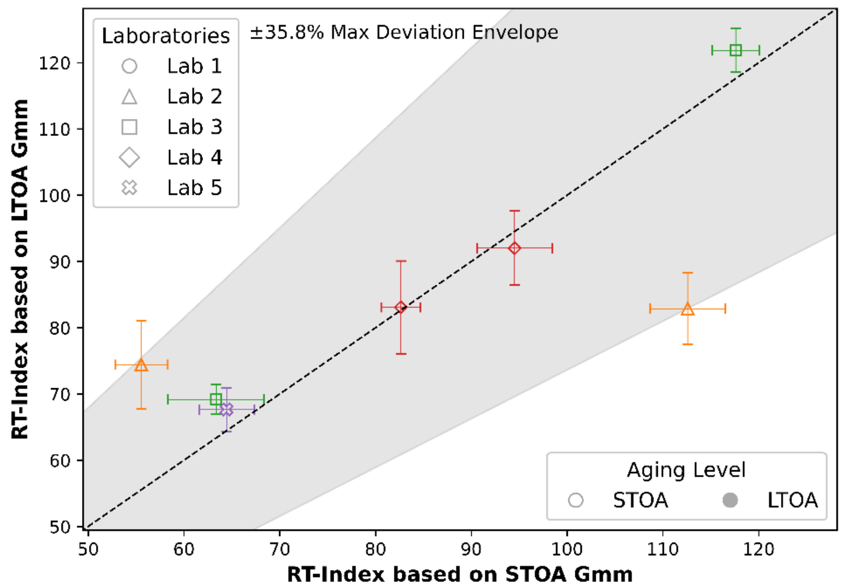


Figure A3.3. Equality plot for RT-indices of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

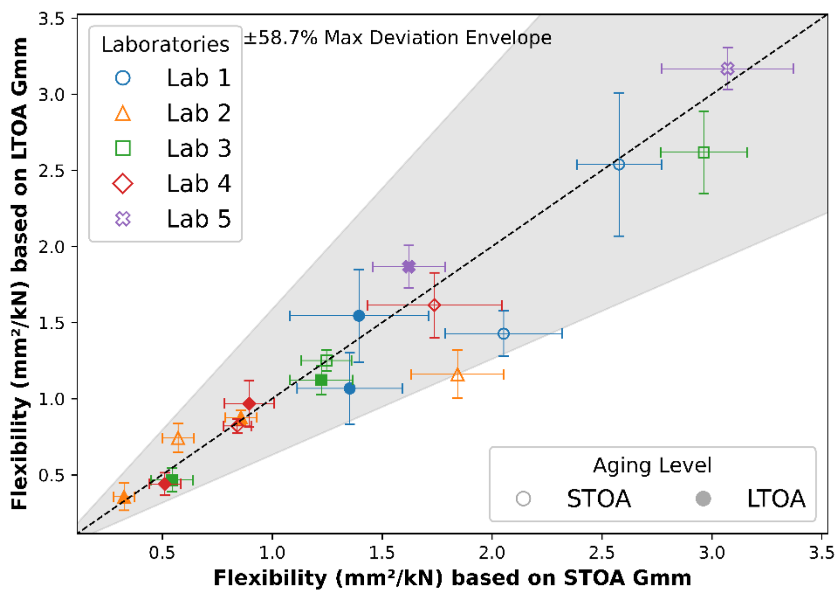


Figure A3.4. Equality plot for flexibility parameter of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

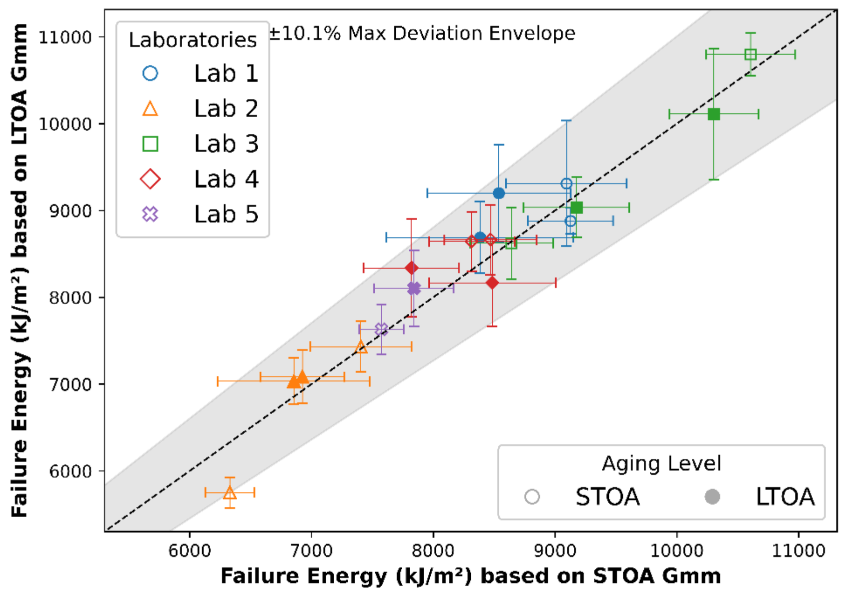


Figure A3.5. Equality plot for failure energy of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

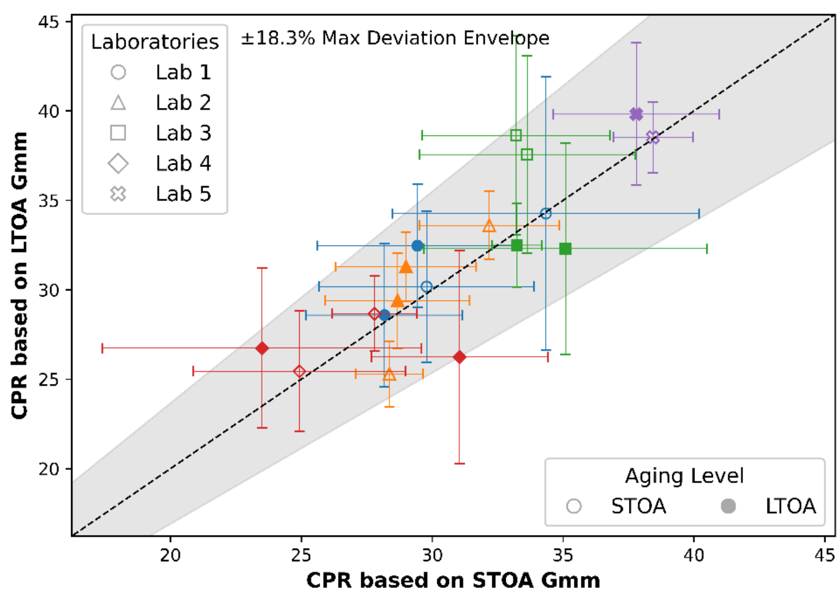


Figure A3.6. Equality plot for CPR of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

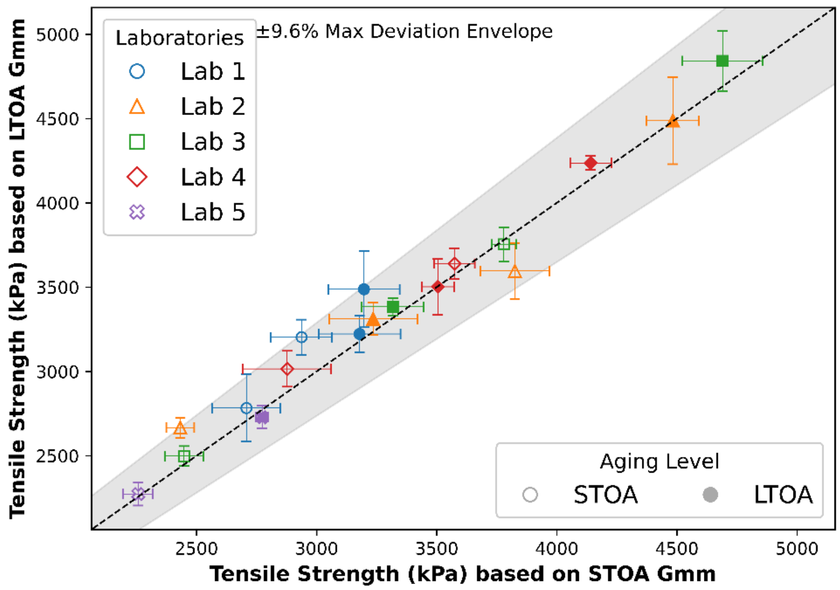


Figure A3.7. Equality plot for tensile strength of specimens compacted based on STOA and LTOA G_{mm} (source: FHWA).

A3.3 p-value Functionality Plots on the Effect of G_{mm} Aging Level

This section includes the p-value functionality plots generated to evaluate the effect of Gmm aging level using different parameters calculated based on IDEAL-CT and IDEAL-RT tests. Each datapoint identifies by the t-statistic and p-value from the Student's t-test on the corresponding sample sets compacted targeting STOA and LTOA G_{mm} aging levels.

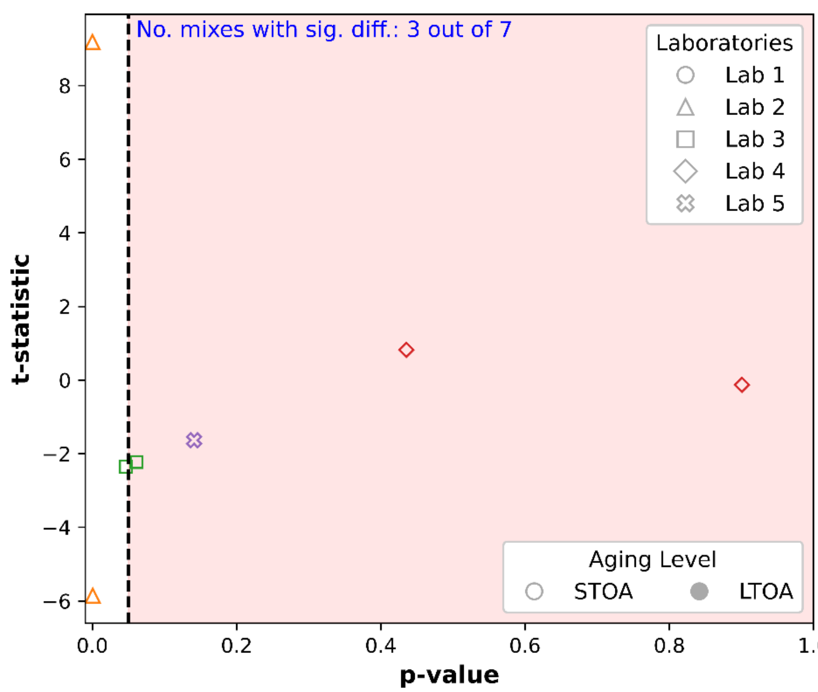


Figure A3.8. p-value functionality plot for RT-index parameter in Task 1 (source: FHWA).

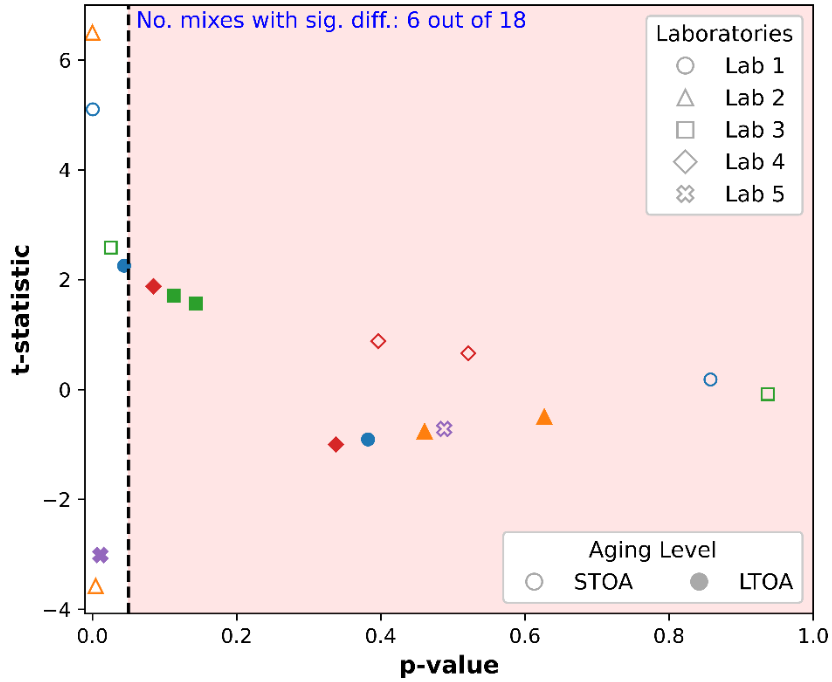


Figure A3.9. p-value functionality plot for flexibility parameter in Task 1 (source: FHWA).

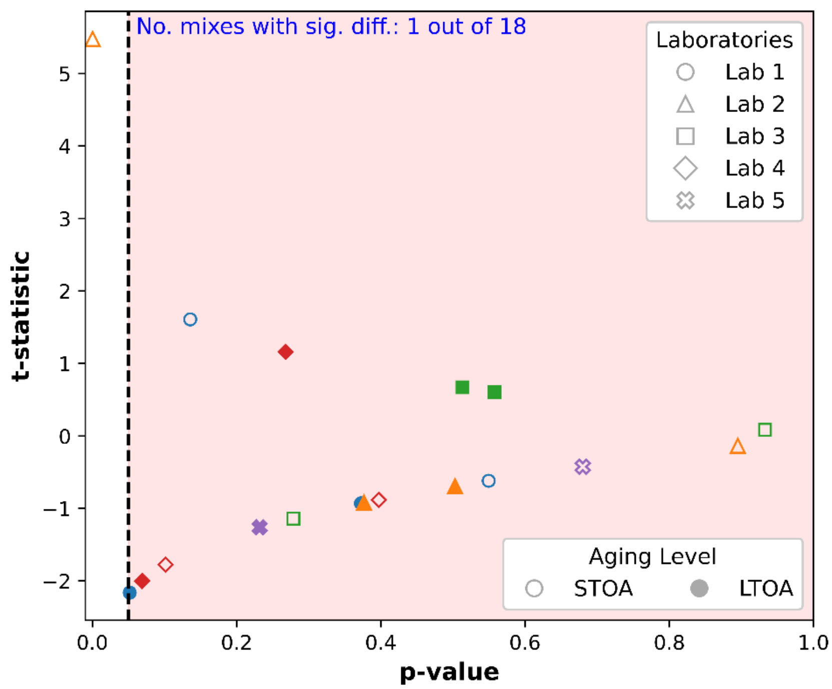


Figure A3.10. p-value functionality plot for failure energy parameter in Task 1 (source: FHWA).

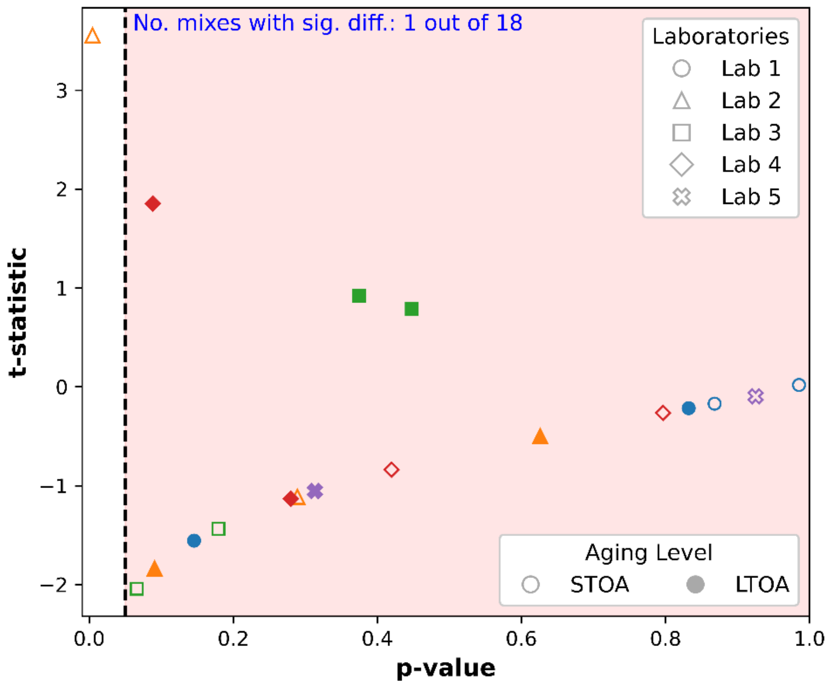


Figure A3.11. p-value functionality plot for CPR parameter in Task 1 (source: FHWA).

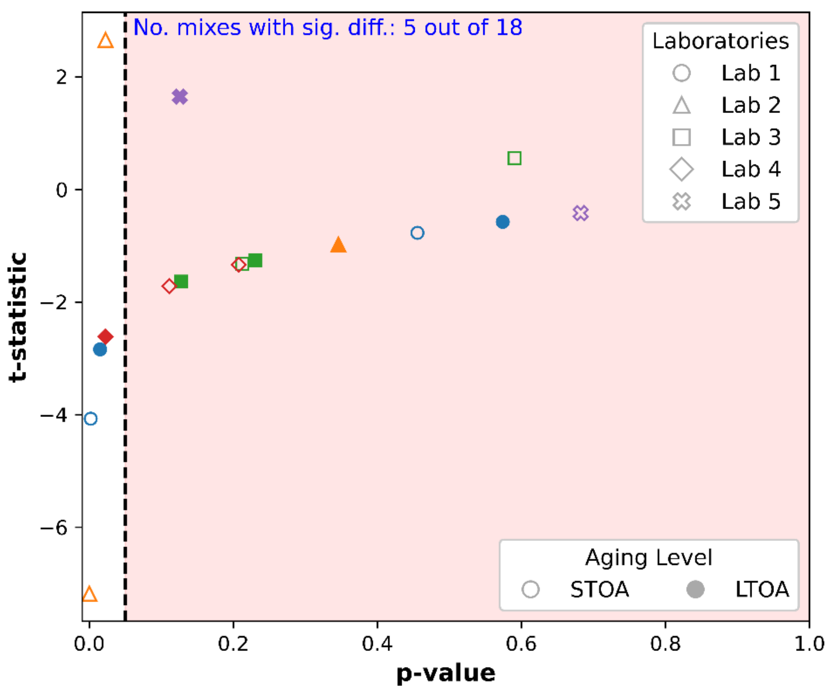


Figure A3.12. p-value functionality plot for tensile strength parameter in Task 1 (source: FHWA).

A3.4 Two-Way ANOVA Results on Effect of G_{mm} Aging Level

This section includes the detailed tabulated results of two-way ANOVA analysis based on different parameters.

Table A3.1. Results of two-way ANOVA based on RT-Index parameter for Task 1.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Gmm Aging Level	1,355.55	1	71.425	0.00
Mixture	46,271.63	8	304.762	0.00
Combined Effect	4,536.21	8	29.877	0.00
Residual	1,328.50	70		

Table A3.2. Results of two-way ANOVA based on flexibility parameter for Task 1.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Gmm Aging Level	0.33	1	1.227	0.27
Mixture	126.74	8	58.706	0.00
Combined Effect	4.76	8	2.204	0.03
Residual	61.00	226		

Table A3.3. Results of two-way ANOVA based on failure energy parameter for Task 1.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Gmm Aging Level	6.77e+5	1	2.708	0.10
Mixture	2.81e+8	8	140.616	0.00
Combined Effect	2.29e+6	8	1.147	0.33
Residual	5.65e+7	226		

Table A3.4. Results of two-way ANOVA based on CPR parameter for Task 1.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Gmm Aging Level	30.16	1	1.766	0.19
Mixture	5144.15	8	37.656	0.00
Combined Effect	85.14	8	0.623	0.76
Residual	3859.17	226		

Table A3.5. Results of two-way ANOVA based on tensile strength parameter for Task 1.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Gmm Aging Level	8.65e+5	1	6.254	0.01
Mixture	7.92e+7	8	71.532	0.00
Combined Effect	1.20e+6	8	1.083	0.38
Residual	3.13e+7	226		

Appendix 4: Comprehensive Analytical Results for Task 2

This appendix provides the supplementary volumetric data, figures, additional analysis and comprehensive statistical outputs supporting the Task 2, which focuses on evaluating the effect of conditioning method on IDEAL-CT and IDEAL-RT performance test results.

A4.1 Air Voids Information

Figure A4.1 shows the distribution of measured air voids for the 13 mixtures evaluated in Task 2. As this figure shows, the laboratory fabrication process yielded highly consistent volumetric properties, with air voids for all compacted test specimens falling within the specified target range of 7.0% ± 0.5%.

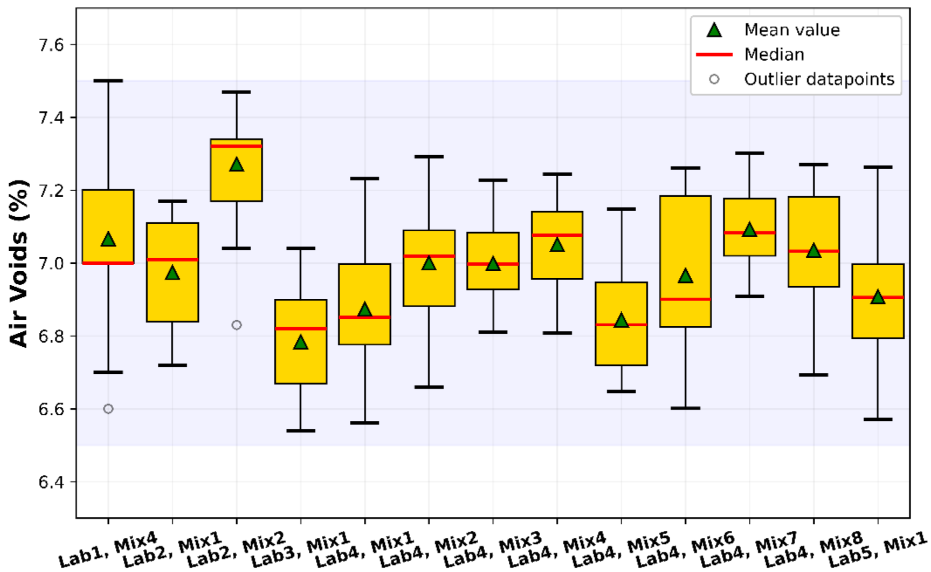


Figure A4.1. Air voids distribution for the specimens used in Task 2 (source: FHWA).

A4.2 Effect of Conditioning Time on Performance Metrics

As a supplementary study outside the primary scope of Task 2, the effect of extended conditioning time during sequential testing on the measured CT- and RT-index values was evaluated. To conduct this analysis, the exact time stamp of testing for each individual specimen was extracted from the raw data files. Within each specific sample set (typically comprising seven replicates for the IDEAL-CT and five for the IDEAL-RT), the initial tested specimen was established as the baseline (time zero). The relative testing delay for all subsequent replicates within that sample set was then calculated in minutes.

While this methodology does not verify the absolute conditioning duration of the baseline specimen prior to its testing, it effectively simulates standard laboratory workflows. In practice, technicians place an entire batch of replicates into the environmental chamber or water bath simultaneously. This inherently causes the final specimens in the queue to undergo longer total conditioning times than the first. Plotting the measured performance indices against this relative testing delay allows for the evaluation of whether this standard sequential testing process inadvertently skews the final results.

The outcomes of this analysis are presented in Figure A4.2 and Figure A4.3 for the CT- and RT-index, respectively, including both the wet and dry conditioning states. Generally, the data indicates no strong correlation between the measured performance indices and the relative conditioning delays. However, a critical boundary condition must be acknowledged: across all five participating laboratories, the sequential testing of an entire sample set was consistently completed within a maximum window of 16 minutes. Therefore, the conclusion that relative conditioning time does not significantly impact mixture performance is strictly limited by this 16-minute operational timeframe.

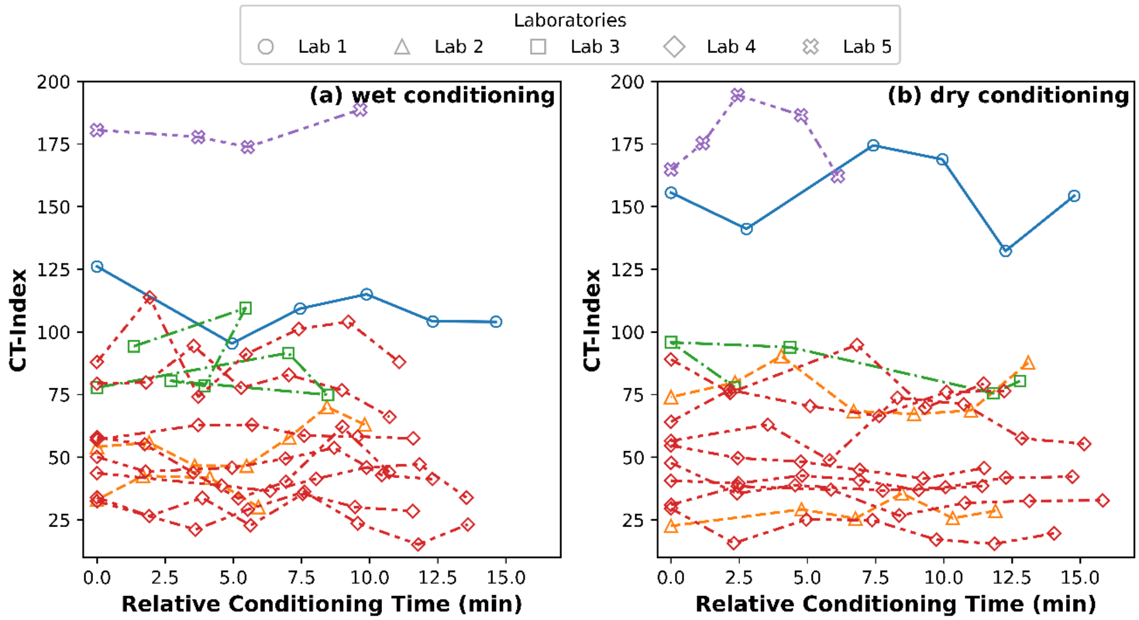


Figure A4.2. Correlation between CT-index and relative conditioning time for: (a) wet conditioning, and (b) dry conditioning (source: FHWA).

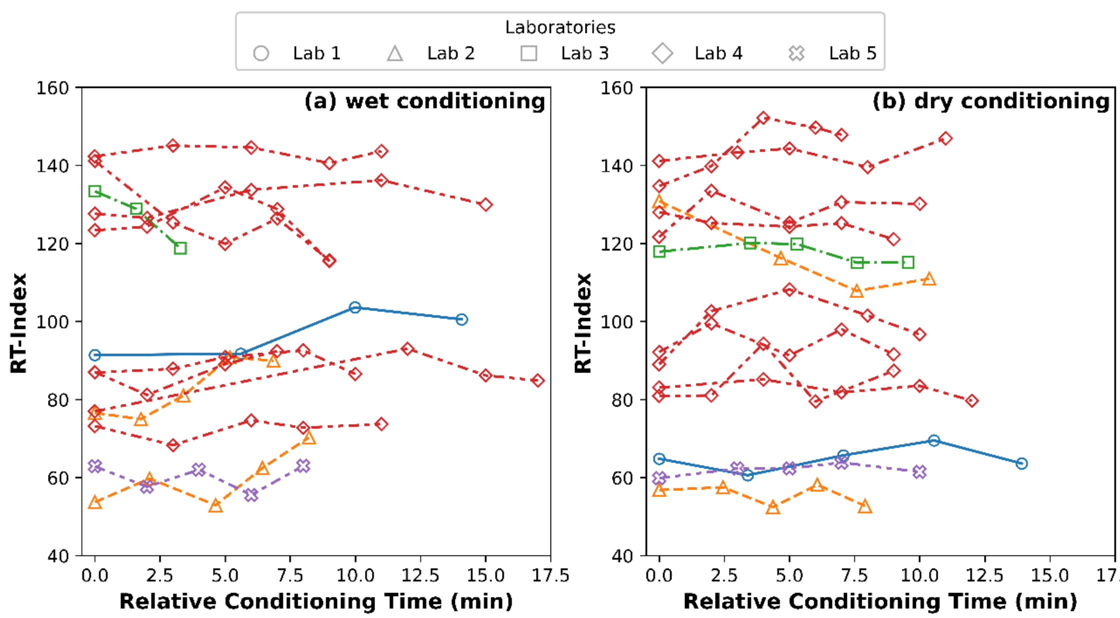


Figure A4.3. Correlation between RT-index and relative conditioning time for (a) wet conditioning, and (b) dry conditioning (source: FHWA).

A4.3 Equality Plots on the Effect of G_{mm} Aging Level

This section includes the equality plots showing the effect of conditioning methods on different parameters calculated for IDEAL-CT and IDEAL-RT tests. Each datapoint on these plots indicates the average parameter value for the sample set conditioning using dry (on horizontal axis) and wet methods (on vertical axis).

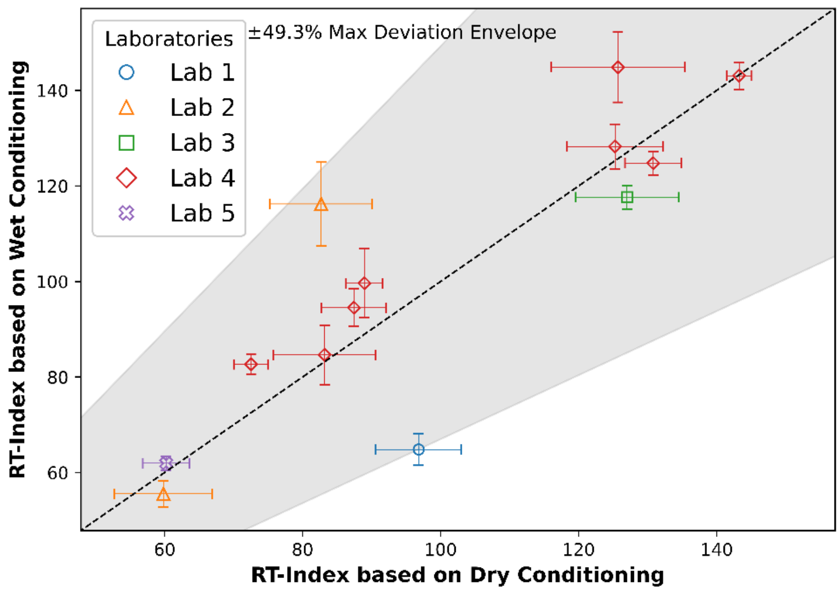


Figure A4.4. Equality plot for RT-index of STOA specimens based on wet and dry conditioning (source: FHWA).

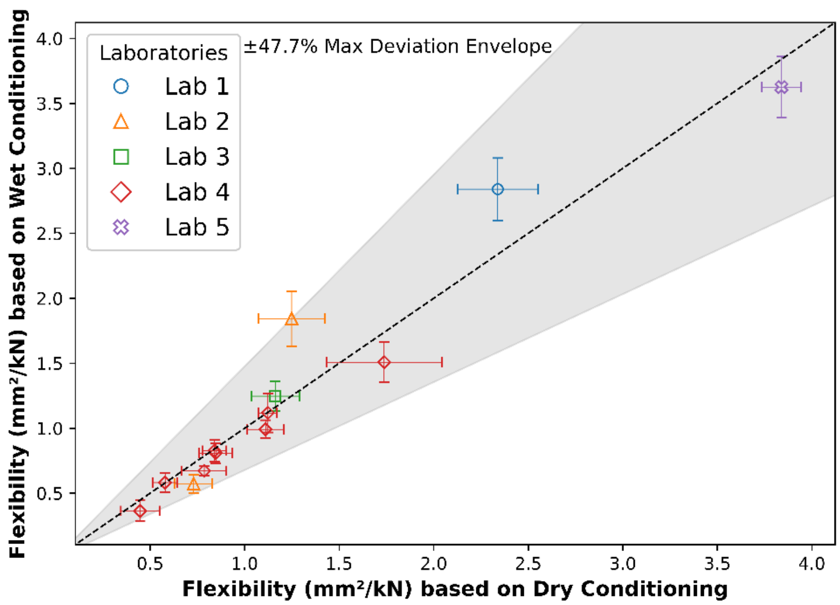


Figure A4.5. Equality plot for flexibility parameter of STOA specimens based on wet and dry conditioning (source: FHWA).

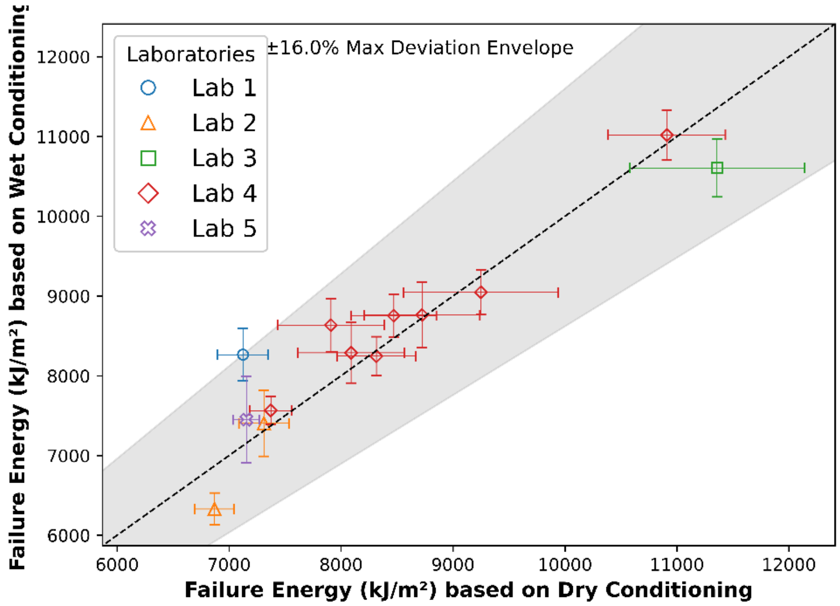


Figure A4.6. Equality plot for failure energy of STOA specimens based on wet and dry conditioning (source: FHWA).

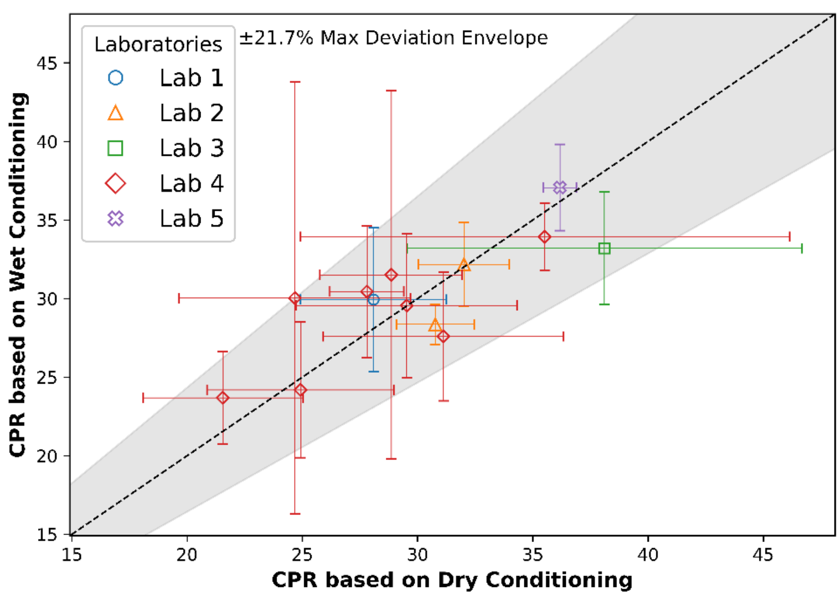


Figure A4.7. Equality plot for CPR of STOA specimens based on wet and dry conditioning (source: FHWA).

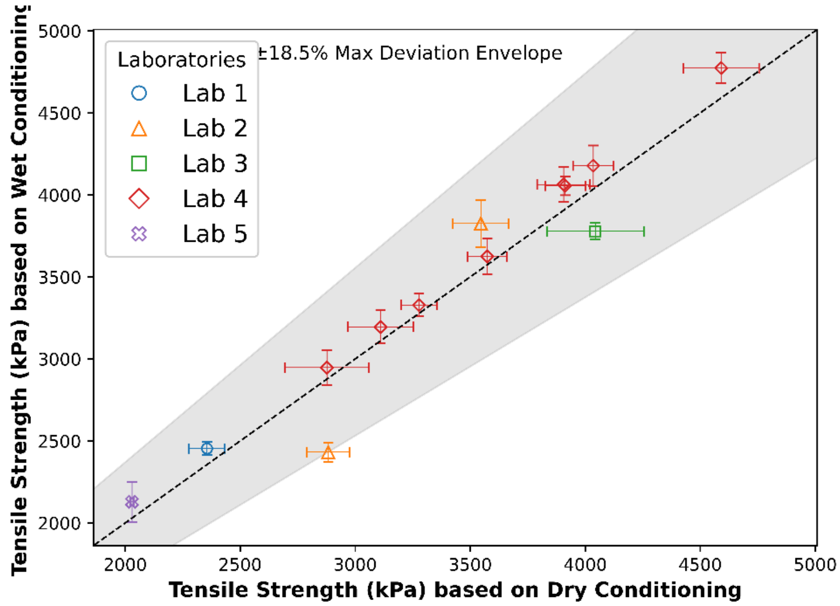


Figure A4.8. Equality plot for tensile strength of specimens based on wet and dry conditioning (source: FHWA).

A4.4 p-value Functionality Plots on the Effect of Conditioning Method

This section includes the p-value functionality plots generated to evaluate the effect of conditioning method using different parameters calculated based on IDEAL-CT and IDEAL-RT tests. Each datapoint identifies by the t-statistic and p-value from the Student's t-test on the corresponding sample sets conditioned using dry and wet methods.

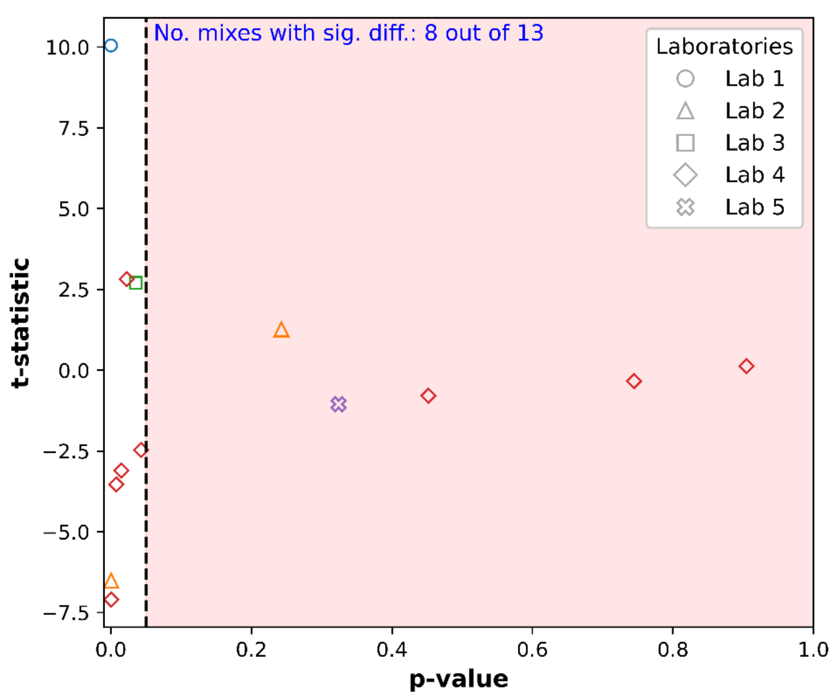


Figure A4.9. p-value functionality plot for RT-index parameter in Task 2 (source: FHWA).

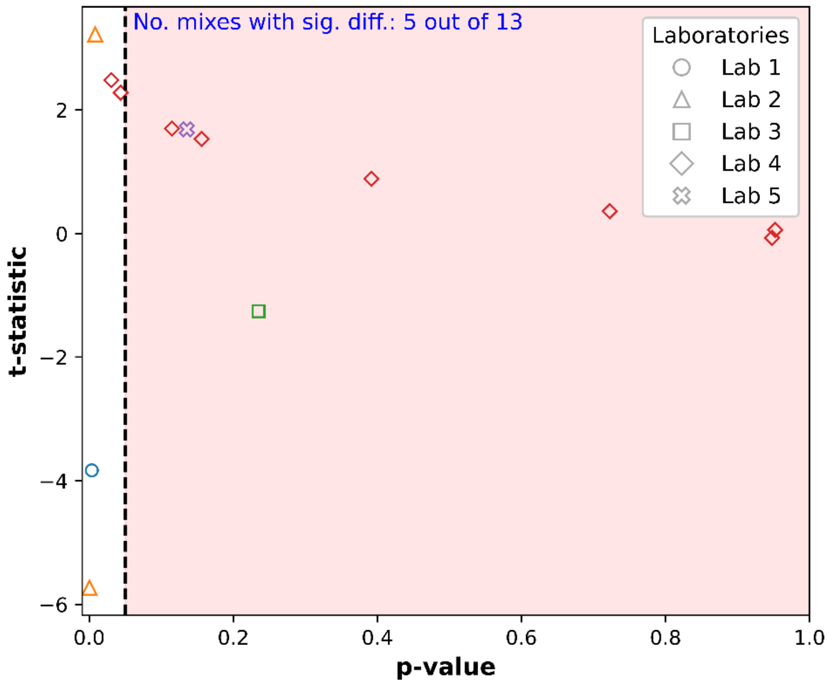


Figure A4.10. p-value functionality plot for flexibility parameter in Task 2 (source: FHWA).

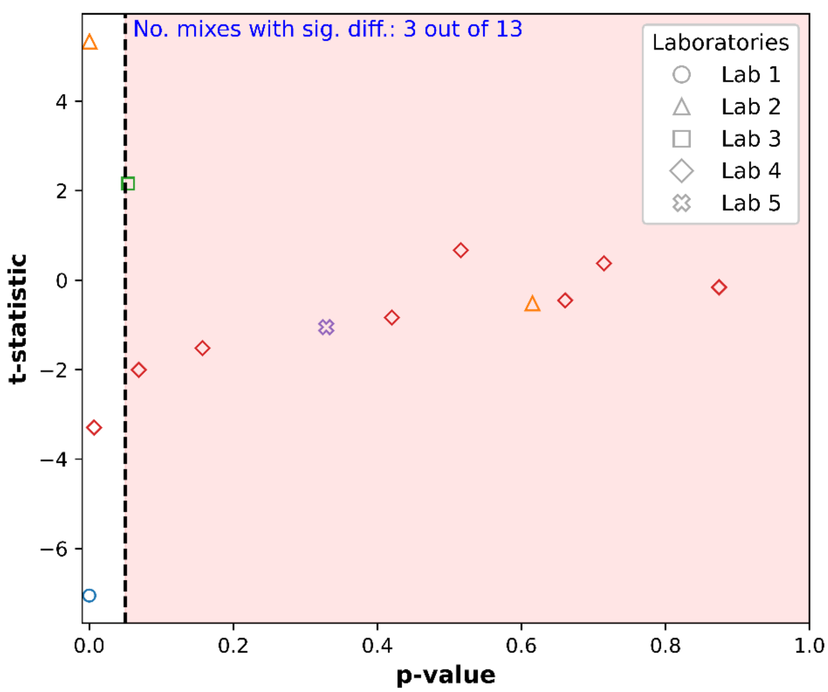


Figure A4.11. p-value functionality plot for failure energy parameter in Task 2 (source: FHWA).

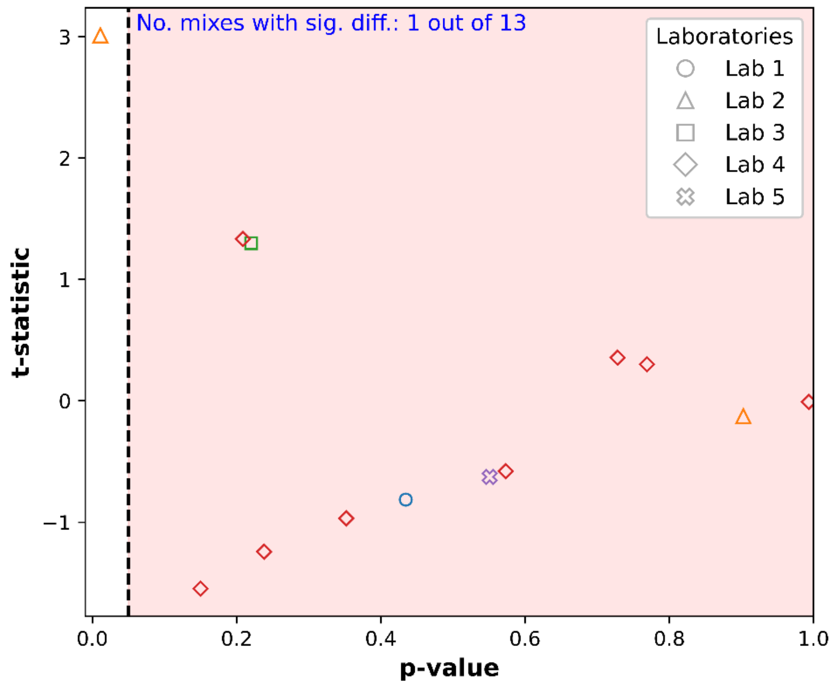


Figure A4.12. p-value functionality plot for CPR parameter in Task 2 (source: FHWA).

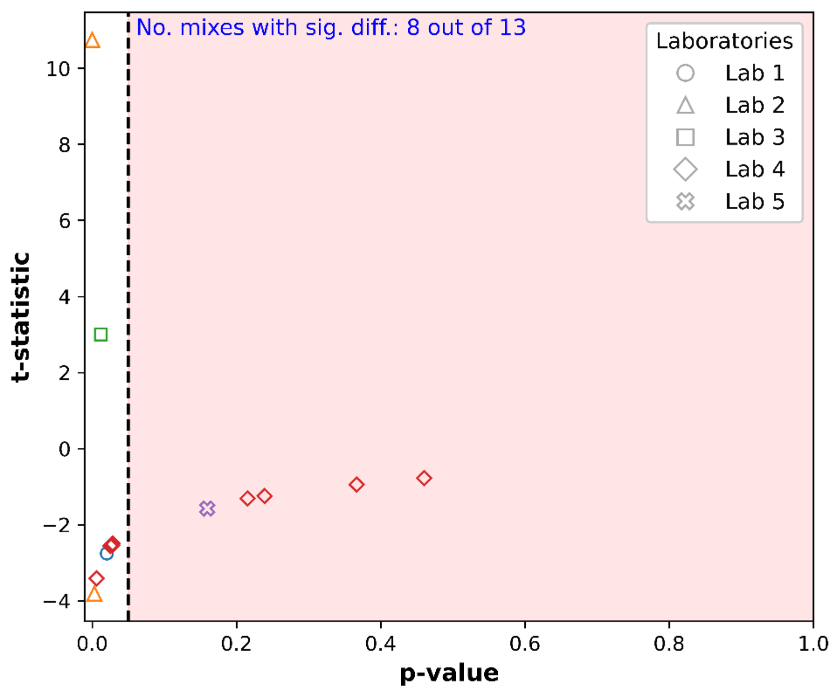


Figure A4.13. p-value functionality plot for tensile strength parameter in Task 2 (source: FHWA).

A4.5 Two-Way ANOVA Results on Effect of Conditioning Method

This section includes the detailed tabulated results of two-way ANOVA analysis based on different parameters.

Table A4.1. Results of two-way ANOVA based on RT-Index parameter for Task 2.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	247.62	1	8.547	0.00
Mixture	267,548.50	12	769.523	0.00
Combined Effect	4,814.95	12	13.849	0.00
Residual	2,897.34	100		

Table A4.2. Results of two-way ANOVA based on flexibility parameter for Task 2.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	0.101	1	5.213	0.02
Mixture	91.504	12	392.396	0.00
Combined Effect	1.260	12	5.402	0.00
Residual	2.740	141		

Table A4.3. Results of two-way ANOVA based on failure energy parameter for Task 2.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	3.85e+5	1	2.408	0.12
Mixture	3.10e+8	12	161.776	0.00
Combined Effect	1.36e+7	12	7.110	0.00
Residual	2.25e+7	141		

Table A4.4. Results of two-way ANOVA based on CPR parameter for Task 2.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	30.070	1	0.935	0.34
Mixture	10,605.684	12	27.494	0.00
Combined Effect	494.060	12	1.280	0.24
Residual	4,532.449	141		

Table A4.5. Results of two-way ANOVA based on tensile strength parameter for Task 2.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	4.69e+4	1	3.737	0.006
Mixture	2.64e+8	12	1,754.051	0.00
Combined Effect	1.96e+6	12	13.002	0.00
Residual	1.77e+6	141		

Appendix 5: Comprehensive Analytical Results for Task 3

This appendix provides the supplementary volumetric data, figures, additional analysis and comprehensive statistical outputs supporting the Task 3, which focuses on evaluating the effect of loose mixture oven heating protocols on IDEAL-CT and IDEAL-RT performance test results.

A5.1 Air Voids Information

Figure A5.1 shows the distribution of measured air voids for the 10 mixtures evaluated in Task 3. As this figure shows, the laboratory fabrication process generally yielded highly consistent volumetric properties, with the air voids for most compacted test specimens falling strictly within the specified target tolerance of 7.0% ± 0.5%. The only exception to this trend was mixture L1-M4, which systematically exhibited slightly lower air void contents, deviating up to 0.5% below the lower bound of the target range.

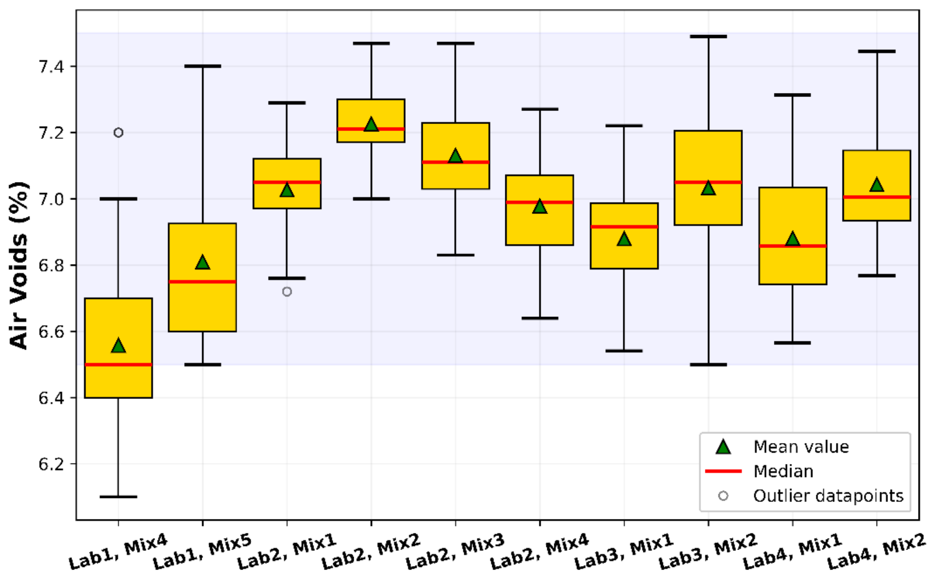


Figure A5.1. Air voids distribution for the specimens used in Task 2 (source: FHWA).

A5.2 Analysis of Oven Heating Durations

All participating laboratories were required to report the specific oven heating durations for each specimen under the assigned heating protocols. These durations are presented as boxplots in Figure A5.2. Analyzing these measured heating times provides a critical quality control check to evaluate how accurately the laboratories executed the intended experimental procedures. A review of the raw data revealed a few discrepancies. First, Laboratory 3 appears to have reported single, expected target times rather than the actual monitored oven heating times for most protocols, which is visually indicated by the flat lines representing zero variance in their corresponding boxplots. Second, for L2-M1 mixture, a severe deviation from the intended protocol was observed for Method A. While the prescribed oven heating time for this method is approximately 180 minutes by task definition, the reported heating duration averaged only 74.5 minutes.

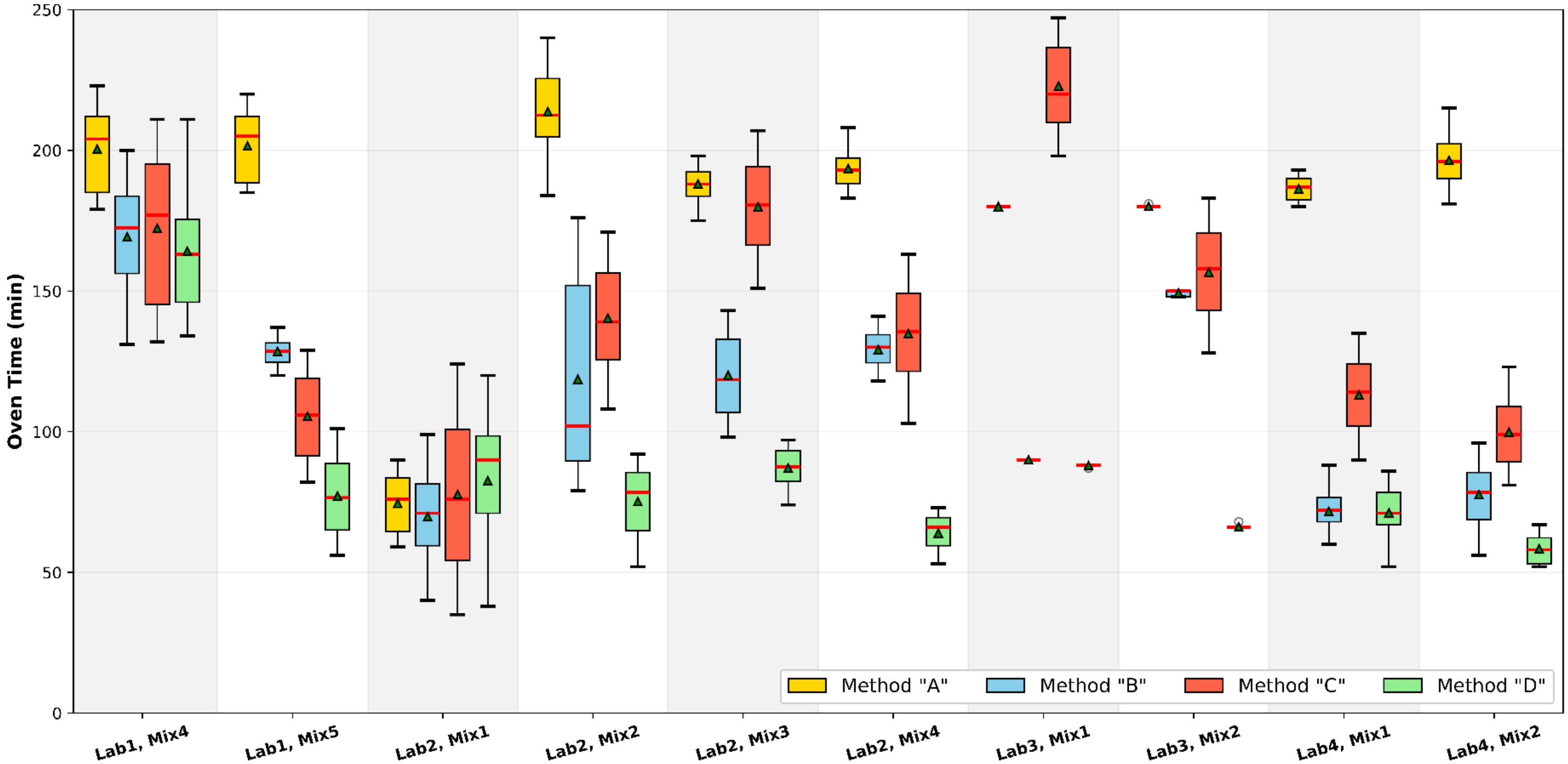


Figure A5.2. Distribution of the oven heating times for given mixture under a specific oven heating protocol (source: FHWA).

Further anomalies were identified regarding the expected logical progression of heating durations between the different methods. Mechanistically, Method C (which requires starting from a cold oven and heating to the target compaction temperature) is expected to require longer heating times than Method B (which utilizes a pre-heated oven). Conversely, for Mixture L1-M5, the average oven times for Methods B and C were reported as 128.5 and 105.5 minutes, respectively. According to the laboratory comments, this anomaly was caused by concurrent testing in the same oven, which increased the overall thermal mass and the number of times the oven doors were opened during Method B testing. Similarly, the heating times for Method D (which targets a temperature 27.8 ℃ (50 ℉) below the compaction temperature) should theoretically be shorter than those for Method B (which targets the full compaction temperature). However, for Mixture L2-M1, the average heating times for Methods B and D were unexpectedly inverted at 69.8 and 82.6 minutes, respectively.

A5.3 Equality Plots on the Effect of Oven Heating Protocol

This section includes the equality plots showing the effect of oven heating protocols on different parameters calculated for IDEAL-CT and IDEAL-RT tests. Each figure is structured as a multi-panel grids. Within each grid, the performance metrics derived from specimen subjected to a specific heating protocol (Methods A, B, C, or D) are plotted against those utilizing Method A, which serves as the experimental reference baseline. Each datapoint on these equality plots indicates the average parameter value for the sample set subjected to oven heating protocol Method A (on horizontal axis) and any other oven heating method (on vertical axis).

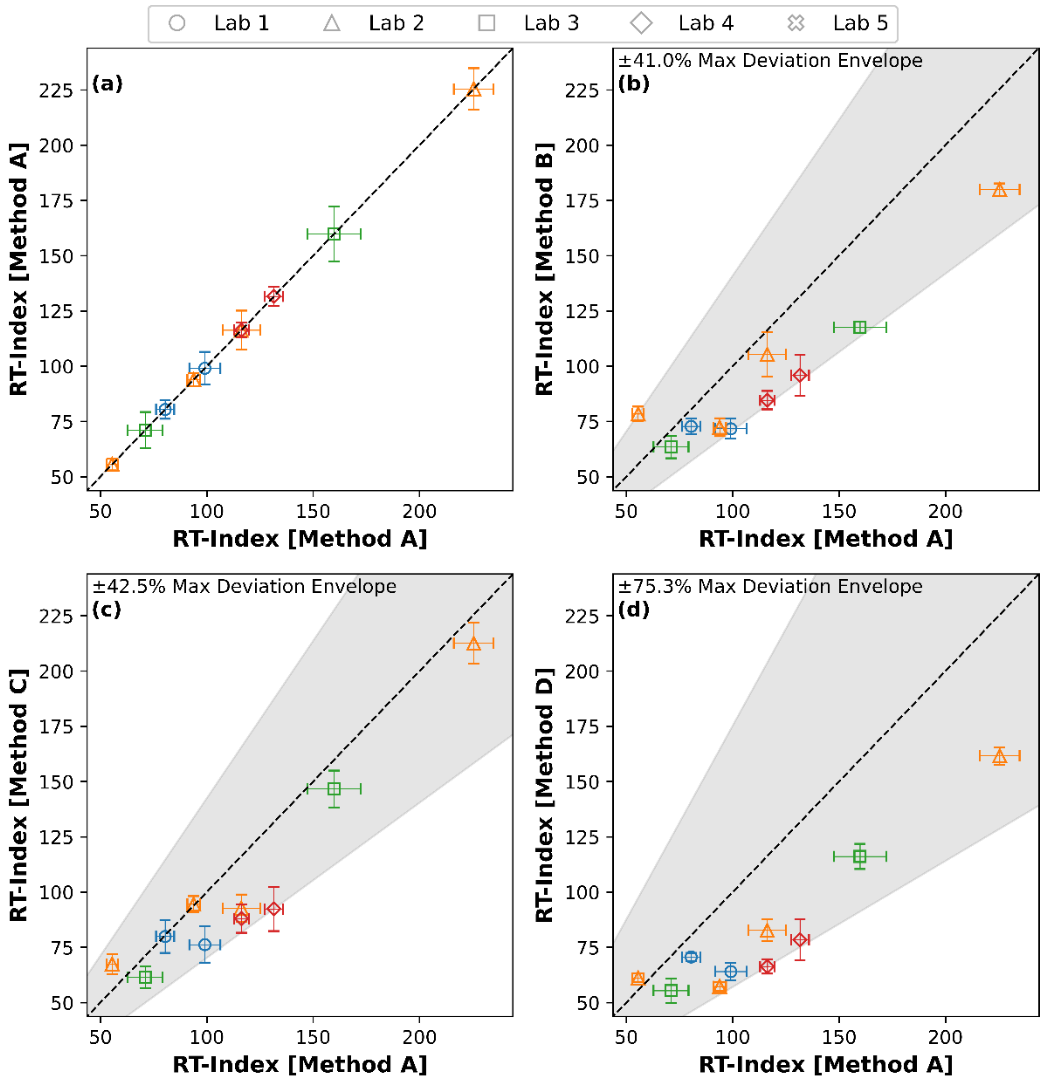


Figure A5.3. Equality plots comparing the RT-indices of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

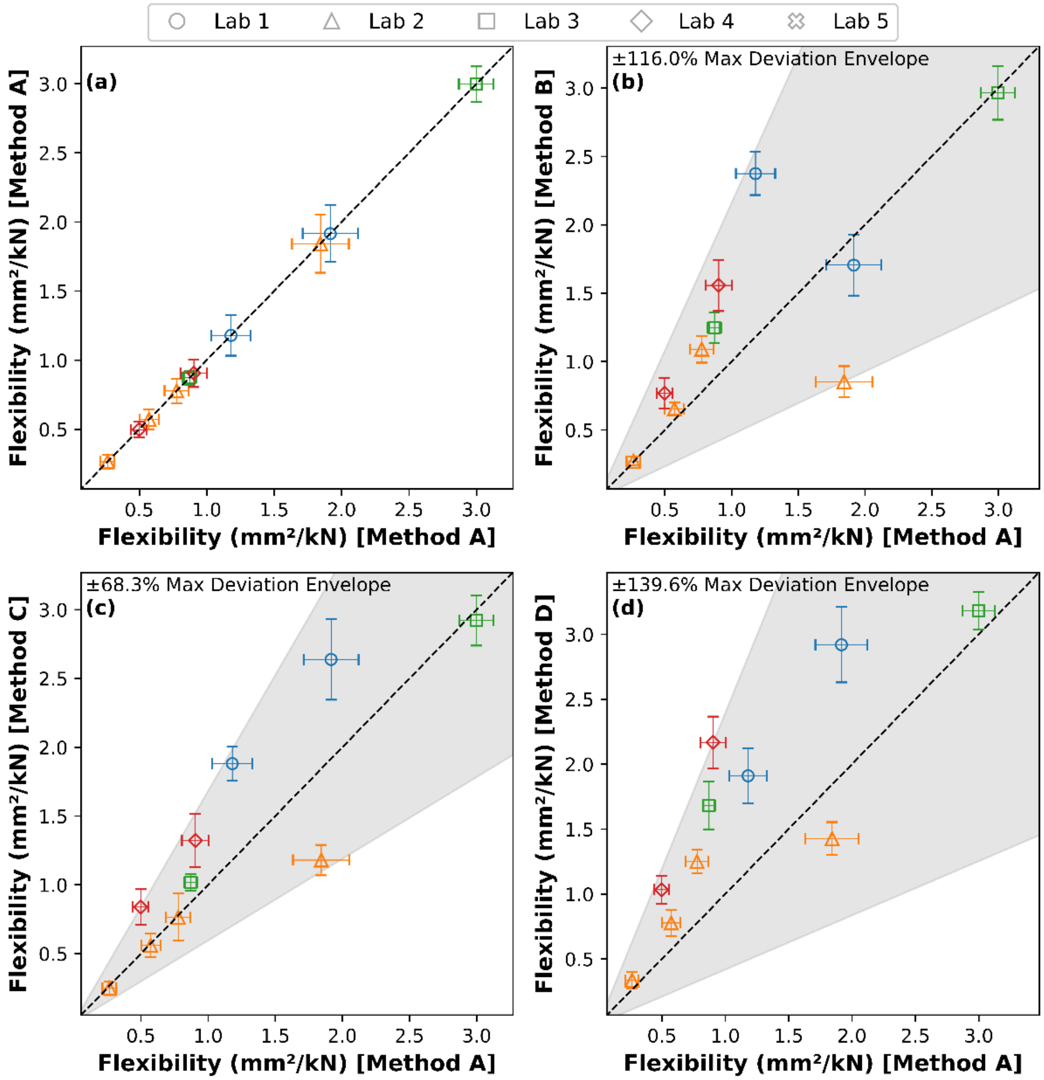


Figure A5.4. Equality plots comparing the flexibility parameter of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

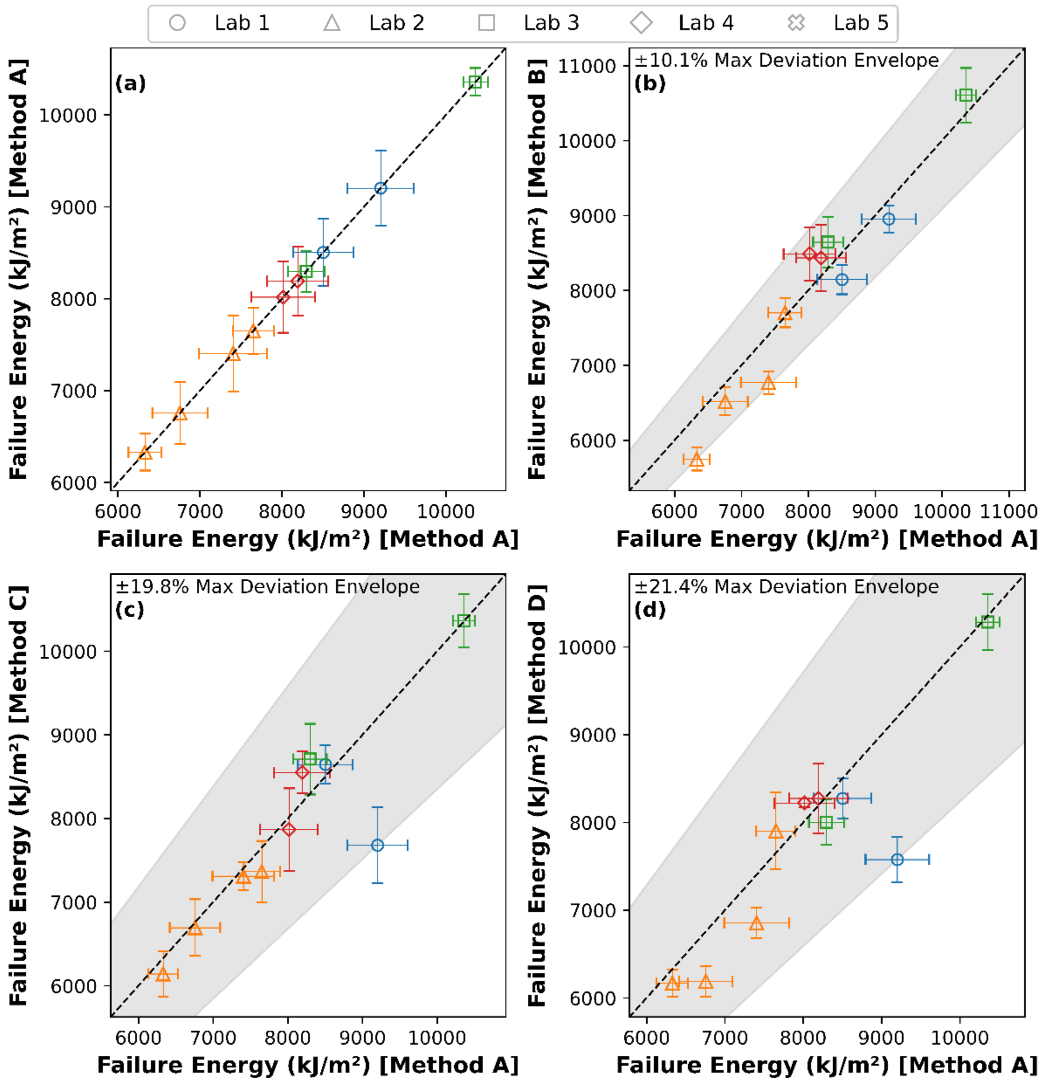


Figure A5.5. Equality plots comparing the failure energy of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

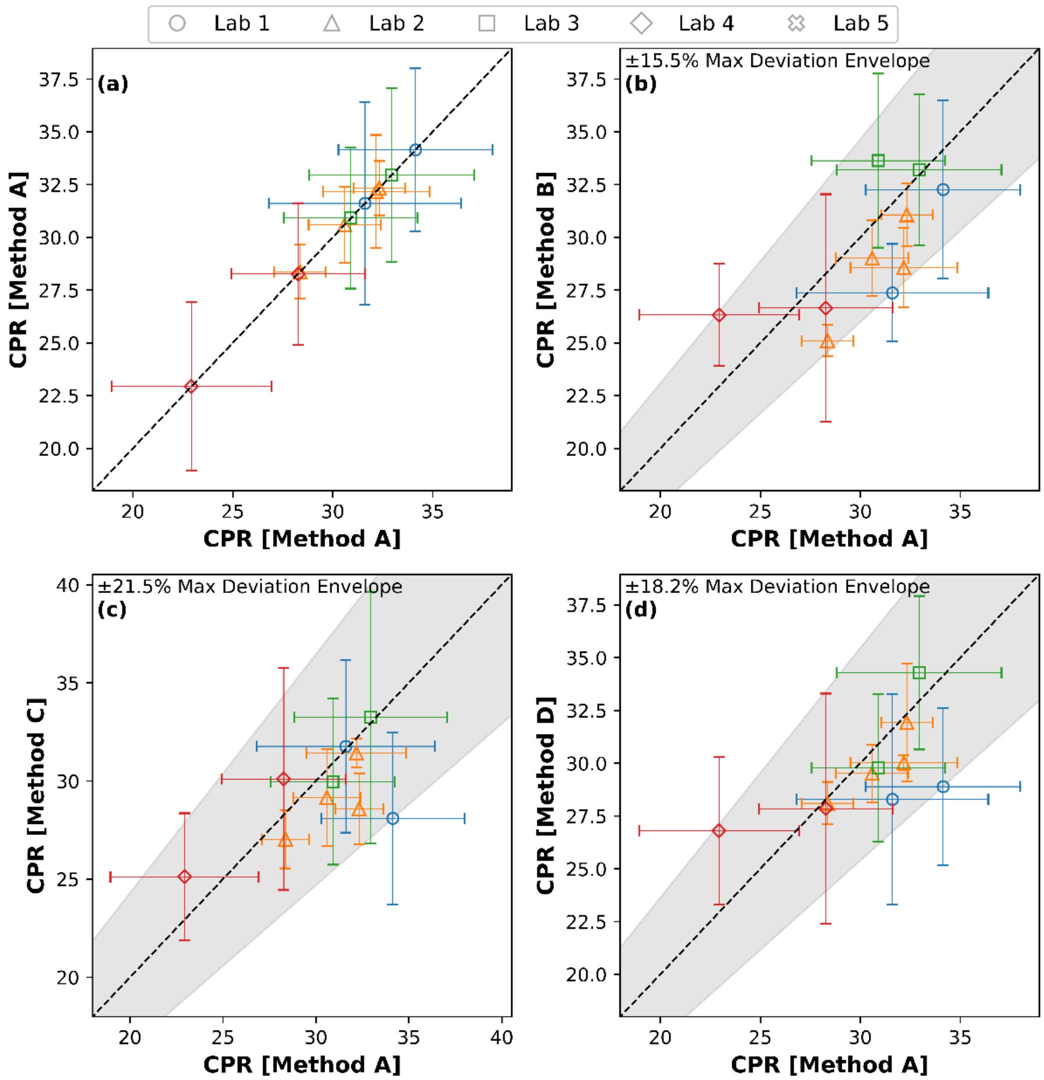


Figure A5.6. Equality plots comparing the CPR of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

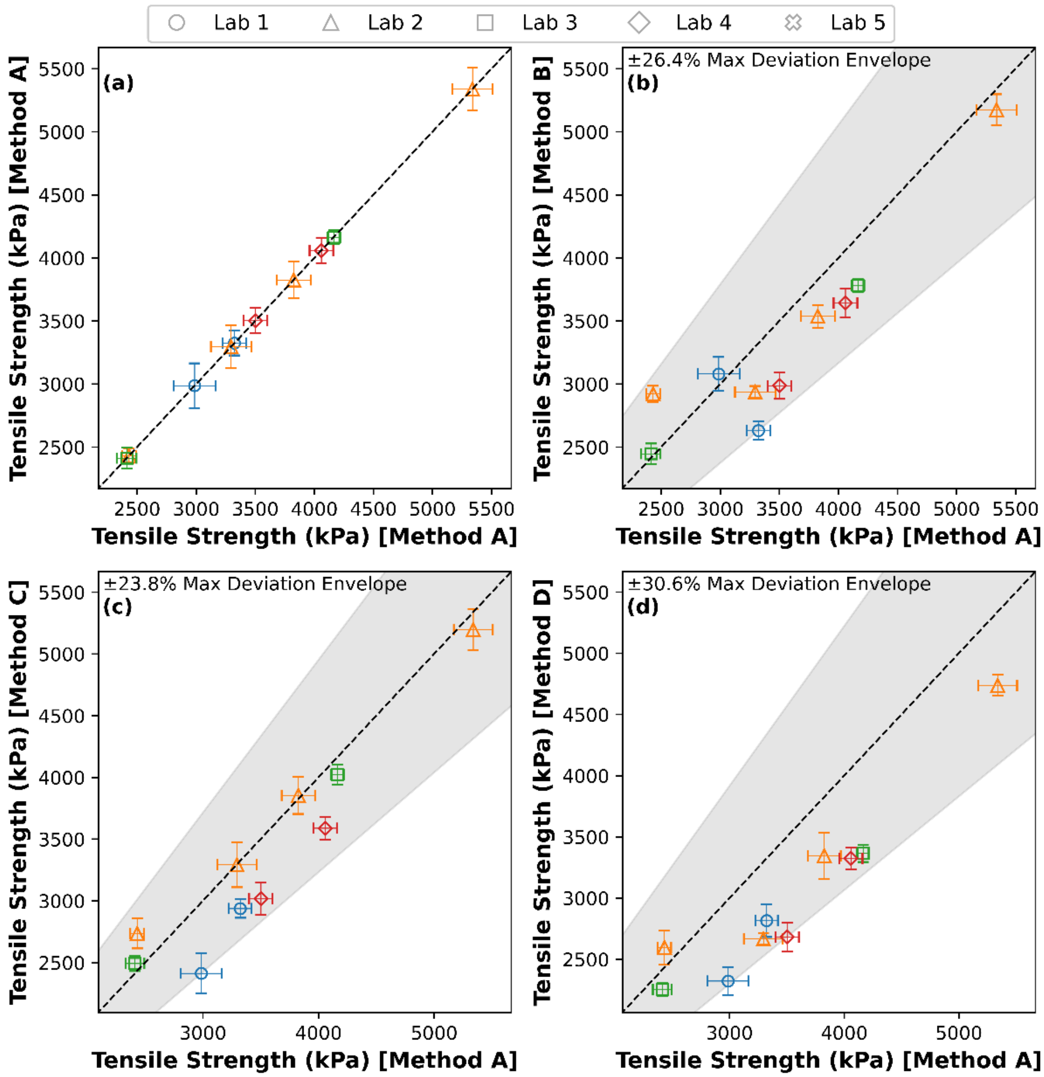


Figure A5.7. Equality plots comparing the tensile strength of specimens subjected to different oven heating protocols against the Method A baseline: (a) Method A vs. Method A, (b) Method B vs. Method A, (c) Method C vs. Method A, and (d) Method D vs. Method A (source: FHWA).

A5.4 One-Way ANOVA Analysis on the Effect of Oven Heating Protocol

To statistically evaluate the specific effect of the four different loose mixture oven heating protocols (Methods A, B, C, and D) on a per-mixture basis, a one-way ANOVA was conducted. Table A5.1 summarizes the calculated F-statistics and corresponding p-values for each of the ten evaluated mixtures across all performance metrics from the IDEAL-CT and IDEAL-RT tests. It is critical to note that the one-way ANOVA serves as an omnibus statistical test. Therefore, a statistically significant result ($p \leq 0.05$) strictly indicates that a significant difference exists between the average performance metrics of at least two of the evaluated heating protocols. However, ANOVA itself does not identify exactly which specific methods significantly differ from one another.

Table A5.1. Summary of one-way ANOVA results evaluating the effect of oven heating protocols across all performance parameters.

Mixture ID	CT-Index		RT-Index		Flexibility Parameter		Failure Energy		CPR		Tensile Strength	
	p-value	**F-stat**	**p-value**	**F-stat**	**p-value**	**F-stat**	**p-value**	**F-stat**	**p-value**	**F-stat**	**p-value**	**F-stat**
L1-M4	✓ 0.00	13.328	✓ 0.01	5.504	✓ 0.00	25.429	✓ 0.00	31.273	✗ 0.12	2.303	✓ 0.00	36.495
L1-M5	✓ 0.00	33.371	✓ 0.00	29.206	✓ 0.00	49.317	✓ 0.02	4.086	✗ 0.24	1.538	✓ 0.00	47.627
L2-M1	✓ 0.00	6.045	✓ 0.00	17.617	✓ 0.00	10.776	✓ 0.00	10.700	✓ 0.00	5.993	✓ 0.00	17.703
L2-M2	✓ 0.00	59.566	✓ 0.00	44.627	✓ 0.00	51.139	✓ 0.00	9.628	✓ 0.00	10.262	✓ 0.00	26.450
L2-M3	✓ 0.00	14.371	✓ 0.00	153.269	✓ 0.00	29.113	✓ 0.00	6.130	✗ 0.42	0.987	✓ 0.00	39.094
L2-M4	✓ 0.01	4.335	✓ 0.00	87.188	✓ 0.03	3.708	✓ 0.04	3.155	✓ 0.01	5.452	✓ 0.00	20.815
L3-M1	✓ 0.00	48.208	✓ 0.00	35.951	✓ 0.00	63.254	✗ 0.26	1.419	✗ 0.95	0.119	✓ 0.00	202.709
L3-M2	✗ 0.91	0.171	✓ 0.01	5.653	✗ 0.06	2.857	✓ 0.00	6.063	✗ 0.31	1.286	✓ 0.00	13.589
L4-M1	✓ 0.00	50.264	✓ 0.00	105.756	✓ 0.00	54.239	✓ 0.04	3.402	✗ 0.22	1.607	✓ 0.00	52.133
L4-M2	✓ 0.00	20.639	✓ 0.00	35.617	✓ 0.00	31.007	✗ 0.30	1.302	✗ 0.64	0.565	✓ 0.00	65.126

A review of the statistical outputs in Table A5.1 reveals several distinct behavioral trends regarding parameter sensitivity. For the RT-index and tensile strength, the effect of the heating protocol is universally significant, yielding p-values well below the 0.05 threshold for all ten evaluated mixtures. Similarly, the composite cracking metrics, like the CT-index and flexibility parameter, exhibit extreme sensitivity to the conditioning protocol, with nine out of ten mixtures exhibiting statistically significant differences (Mixture L3-M2 being the only exception for both parameters). The failure energy parameter also displays a high degree of sensitivity, showing significant variance across eight of the ten mixtures.

Conversely, the CPR stands out as the only metric largely unaffected by the variations in the oven heating protocols. For the CPR parameter, seven out of the ten mixtures yielded insignificant p-values ($p > 0.05$). This outcome reinforces the conclusions drawn in Tasks 1 and 2, further confirming that this specific foundational property exhibited lower statistical sensitivity and remains robust against procedural changes during laboratory fabrication compared to the mathematically derived composite indices.

A5.5. Post-Hoc Analysis on Effect of Oven Heating Protocol

Based on the omnibus one-way ANOVA results, the average performance metrics for most mixtures were significantly affected by the different oven heating protocols. To explicitly determine which specific oven heating protocols resulted in significantly different performance metrics for a given mixture and parameter, a Games-Howell post-hoc analysis was conducted. The results of these pairwise comparisons are provided in Table A5.2 through Table A5.11. In these tables, all possible combinations of the oven heating protocols are compared against one another, and their corresponding p-values are reported. A p-value equal to or less than 0.05 indicates that the mean performance metric values are significantly different. To facilitate rapid visual interpretation, the “Is Diff?” columns utilize a check mark (✓) to denote a statistically significant difference and a cross symbol (✗) to indicate no significant difference.

Table A5.2. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L1-M4.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*
A	B	0.29	✗	0.06	✗	0.50	✗	0.69	✗	0.90	✗	0.82	✗
A	C	0.14	✗	1.00	✗	0.01	✓	0.01	✓	0.21	✗	0.01	✓
A	D	0.03	✓	0.02	✓	0.00	✓	0.00	✓	0.24	✗	0.01	✓
B	C	0.01	✓	0.31	✗	0.00	✓	0.01	✓	0.46	✗	0.00	✓
B	D	0.00	✓	0.66	✗	0.00	✓	0.00	✓	0.54	✗	0.00	✓
C	D	0.40	✗	0.15	✗	0.43	✗	0.97	✗	0.99	✗	0.72	✗

Table A5.3. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L1-M5.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*
A	B	0.00	✓	0.00	✓	0.00	✓	0.24	✗	0.30	✗	0.00	✓
A	C	0.00	✓	0.01	✓	0.00	✓	0.85	✗	1.00	✗	0.00	✓
A	D	0.00	✓	0.00	✓	0.00	✓	0.57	✗	0.63	✗	0.00	✓
B	C	0.03	✓	0.71	✗	0.00	✓	0.02	✓	0.23	✗	0.00	✓
B	D	0.02	✓	0.08	✗	0.01	✓	0.72	✗	0.97	✗	0.05	✗
C	D	0.95	✗	0.09	✗	0.99	✗	0.06	✗	0.56	✗	0.24	✗

Table A5.4. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L2-M1.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*
A	B	0.93	✗	0.33	✗	0.16	✗	0.04	✓	0.09	✗	0.01	✓
A	C	0.97	✗	0.01	✓	0.99	✗	0.95	✗	0.91	✗	0.98	✗
A	D	0.05	✗	0.00	✓	0.01	✓	0.08	✗	0.31	✗	0.00	✓
B	C	0.63	✗	0.17	✗	0.11	✗	0.00	✓	0.04	✓	0.00	✓
B	D	0.04	✓	0.02	✓	0.07	✗	0.79	✗	0.34	✗	0.16	✗
C	D	0.01	✓	0.09	✗	0.00	✓	0.00	✓	0.01	✓	0.00	✓

Table A5.5. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L2-M2.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.00	✓	0.00	✓	0.00	✓	0.00	✓	0.00	✓	0.00	✓
A	C	0.00	✓	0.01	✓	0.00	✓	0.53	✗	0.37	✗	0.00	✓
A	D	0.01	✓	0.03	✓	0.01	✓	0.37	✗	0.97	✗	0.08	✗
B	C	0.00	✓	0.01	✓	0.00	✓	0.06	✗	0.09	✗	0.05	✓
B	D	0.00	✓	0.00	✓	0.00	✓	0.00	✓	0.00	✓	0.00	✓
C	D	0.00	✓	0.11	✗	0.01	✓	1.00	✗	0.48	✗	0.26	✗

Table A5.6. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L2-M3.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.00	✓	0.00	✓	0.00	✓	0.41	✗	0.39	✗	0.00	✓
A	C	1.00	✗	0.99	✗	1.00	✗	0.99	✗	0.62	✗	1.00	✗
A	D	0.00	✓	0.00	✓	0.00	✓	0.01	✓	0.60	✗	0.00	✓
B	C	0.04	✓	0.00	✓	0.01	✓	0.64	✗	1.00	✗	0.01	✓
B	D	0.45	✗	0.00	✓	0.04	✓	0.02	✓	0.93	✗	0.00	✓
C	D	0.01	✓	0.00	✓	0.00	✓	0.03	✓	0.99	✗	0.00	✓

Table A5.7. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L2-M4.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.97	✗	0.00	✓	0.99	✗	0.98	✗	0.37	✗	0.22	✗
A	C	0.83	✗	0.20	✗	0.92	✗	0.37	✗	0.00	✓	0.42	✗
A	D	0.19	✗	0.00	✓	0.19	✗	0.62	✗	0.99	✗	0.00	✓
B	C	0.57	✗	0.00	✓	0.71	✗	0.21	✗	0.07	✗	0.99	✗
B	D	0.24	✗	0.00	✓	0.21	✗	0.72	✗	0.90	✗	0.00	✓
C	D	0.09	✗	0.00	✓	0.08	✗	0.14	✗	0.13	✗	0.00	✓

Table A5.8. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L3-M1.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.00	✓	0.00	✓	0.00	✓	0.47	✗	1.00	✗	0.00	✓
A	C	0.00	✓	0.28	✗	0.00	✓	1.00	✗	1.00	✗	0.02	✓
A	D	0.00	✓	0.00	✓	0.00	✓	0.94	✗	0.91	✗	0.00	✓
B	C	0.02	✓	0.00	✓	0.01	✓	0.62	✗	1.00	✗	0.00	✓
B	D	0.01	✓	0.95	✗	0.00	✓	0.38	✗	0.95	✗	0.00	✓
C	D	0.00	✓	0.00	✓	0.00	✓	0.97	✗	0.98	✗	0.00	✓

Table A5.9. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L3-M2.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.90	✗	0.36	✗	0.98	✗	0.25	✗	0.64	✗	0.89	✗
A	C	0.95	✗	0.21	✗	0.84	✗	0.24	✗	0.98	✗	0.35	✗
A	D	0.91	✗	0.04	✓	0.18	✗	0.25	✗	0.94	✗	0.03	✓
B	C	1.00	✗	0.92	✗	0.98	✗	0.99	✗	0.46	✗	0.71	✗
B	D	1.00	✗	0.15	✗	0.19	✗	0.02	✓	0.35	✗	0.00	✓
C	D	1.00	✗	0.33	✗	0.08	✗	0.03	✓	1.00	✗	0.00	✓

Table A5.10. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L4-M1.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.00	✓	0.00	✓	0.00	✓	0.19	✗	0.35	✗	0.00	✓
A	C	0.02	✓	0.00	✓	0.00	✓	0.93	✗	0.71	✗	0.00	✓
A	D	0.00	✓	0.00	✓	0.00	✓	0.60	✗	0.34	✗	0.00	✓
B	C	0.06	✗	0.77	✗	0.18	✗	0.09	✗	0.87	✗	0.96	✗
B	D	0.00	✓	0.00	✓	0.00	✓	0.36	✗	0.99	✗	0.00	✓
C	D	0.00	✓	0.00	✓	0.00	✓	0.32	✗	0.81	✗	0.00	✓

Table A5.11. Games-Howell pairwise comparison of performance metrics across oven heating protocols for mixture L4-M2.

Group 1	Group 2	CT-Index		RT-Index		Flexibility Param.		Failure Energy		CPR		Tensile Strength	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
A	B	0.00	✓	0.00	✓	0.00	✓	0.69	✗	0.90	✗	0.00	✓
A	C	0.00	✓	0.00	✓	0.00	✓	0.21	✗	0.88	✗	0.00	✓
A	D	0.00	✓	0.00	✓	0.00	✓	0.98	✗	1.00	✗	0.00	✓
B	C	0.72	✗	0.94	✗	0.71	✗	0.93	✗	0.66	✗	0.74	✗
B	D	0.03	✓	0.07	✗	0.00	✓	0.89	✗	0.98	✗	0.00	✓
C	D	0.21	✗	0.18	✗	0.04	✓	0.44	✗	0.87	✗	0.00	✓

A5.6 Two-Way ANOVA Results on Effect of Oven Heating Protocol

This section includes the detailed tabulated results of two-way ANOVA analysis based on different parameters.

Table A5.12. Results of two-way ANOVA based on RT-Index parameter for Task 3.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	34,154.14	3	293.703	0.00
Mixture	369,391.71	9	1,058.842	0.00
Combined Effect	36,229.73	27	34.617	0.00
Residual	6,202.02	160		

Table A5.13. Results of two-way ANOVA based on flexibility parameter for Task 3.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	7.61	3	122.917	0.00
Mixture	178.51	9	960.550	0.00
Combined Effect	11.56	27	20.731	0.00
Residual	4.40	213		

Table A5.14. Results of two-way ANOVA based on failure energy parameter for Task 3.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	3.84e+6	3	13.922	0.00
Mixture	5.75e+8	9	693.704	0.00
Combined Effect	2.00e+7	27	8.028	0.00
Residual	1.96e+7	213		

Table A5.15. Results of two-way ANOVA based on CPR parameter for Task 3.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	468.17	3	13.424	0.00
Mixture	2,424.22	9	23.169	0.00
Combined Effect	781.53	27	2.490	0.00
Residual	2476.26	213		

Table A5.16. Results of two-way ANOVA based on tensile strength parameter for Task 3.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	1.16e+7	3	289.279	0.00
Mixture	1.78e+8	9	1484.432	0.00
Combined Effect	1.04e+7	27	28.877	0.00
Residual	2.84e+6	213		

Appendix 6: Comprehensive Analytical Results for Task 4

This appendix provides the supplementary volumetric data, figures, additional analysis and comprehensive statistical outputs supporting the Task 4, which focuses on evaluating the effect of specimen trimming method on HWTT performance test results.

A6.1 Air Voids Information

Figure A6.1 shows the distribution of measured air voids for the seven plant-produced mixtures evaluated in Task 4, categorized by their respective specimen trimming protocols (Methods A, B, and C). As visually delineated by the shaded background region in the plot, the laboratory fabrication process yielded highly consistent volumetric properties, with most compacted test specimens successfully falling within the prescribed target tolerance of 7.0% ± 0.5%.

Additionally, a review of this data indicates that the choice of trimming method did not introduce a systematic or universal bias into the final specimen air voids. For example, while specimens subjected to double-sided trimming (Method C) exhibited lower average air voids compared to their directly compacted counterparts (Method A) for mixtures L2-M2, L4-M1, and L5-M1, this reduction was not observed across the remaining four mixtures.

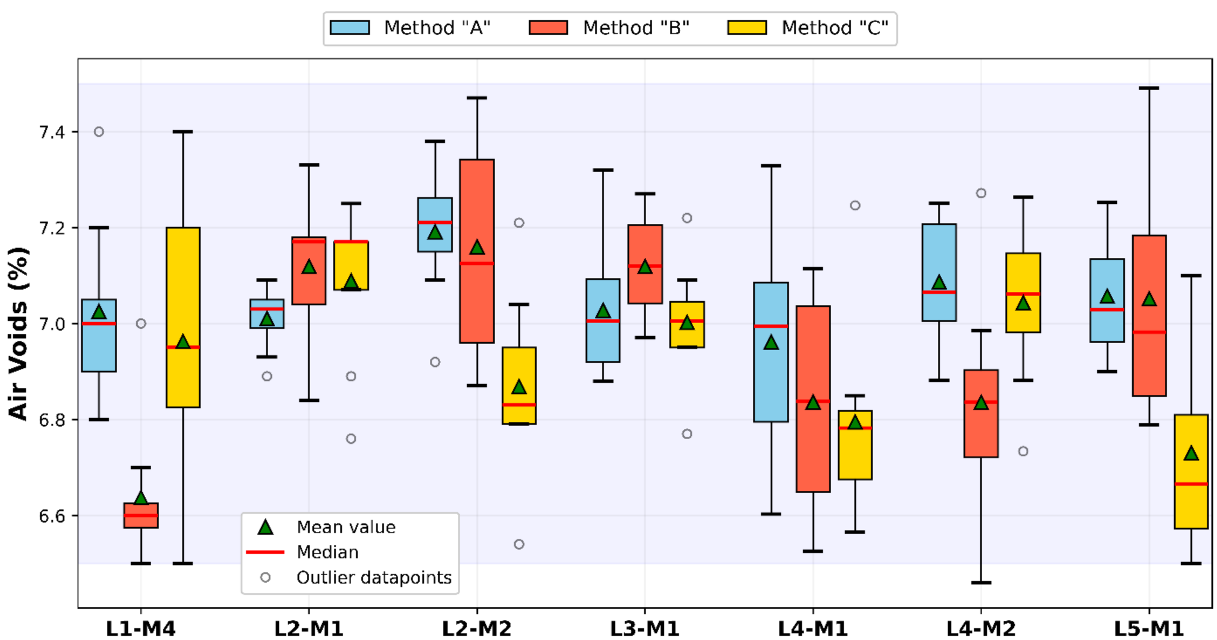


Figure A6.1. Effect of specimen trimming method on the final compacted air voids of Task 4 mixtures (source: FHWA).

A6.2 Performance Metrics from HWTT Results

This section includes the bar plots of the calculated performance metrics from the HWTT rutting curves across different mixtures and trimming methods. These parameters were calculated using the AutoHWTT software based on fitting the 2PP model, including total rut depth (TRD), corrected rut depth at 20,000 passes (CRD20), stripping number (SN), Stripping Inflection Point (SIP) and adjusted SIP.

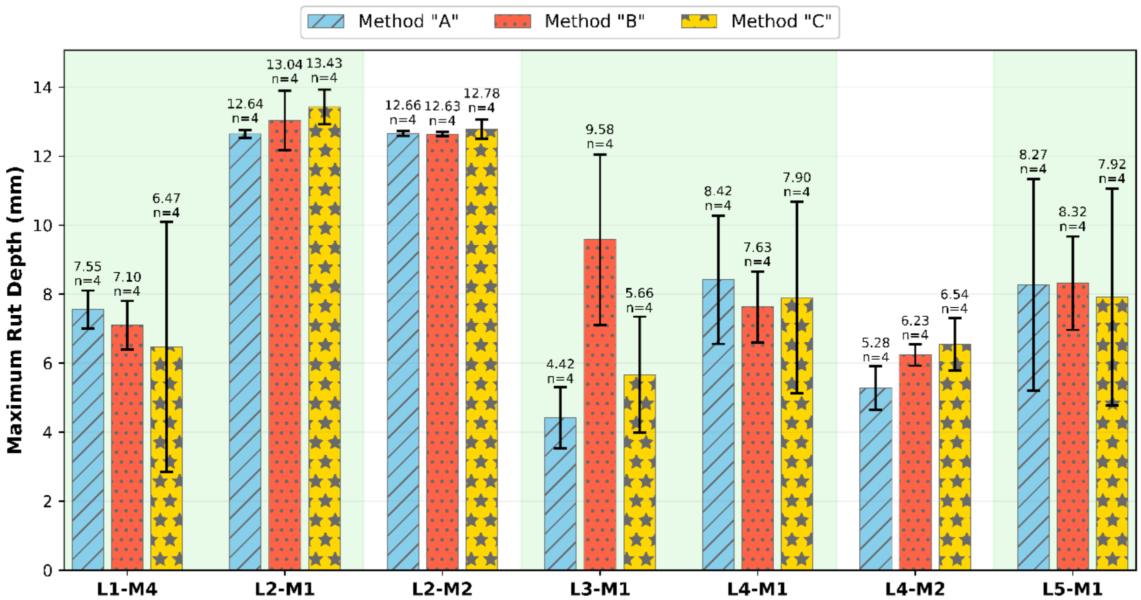


Figure A6.2. Summary of maximum rut depths for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

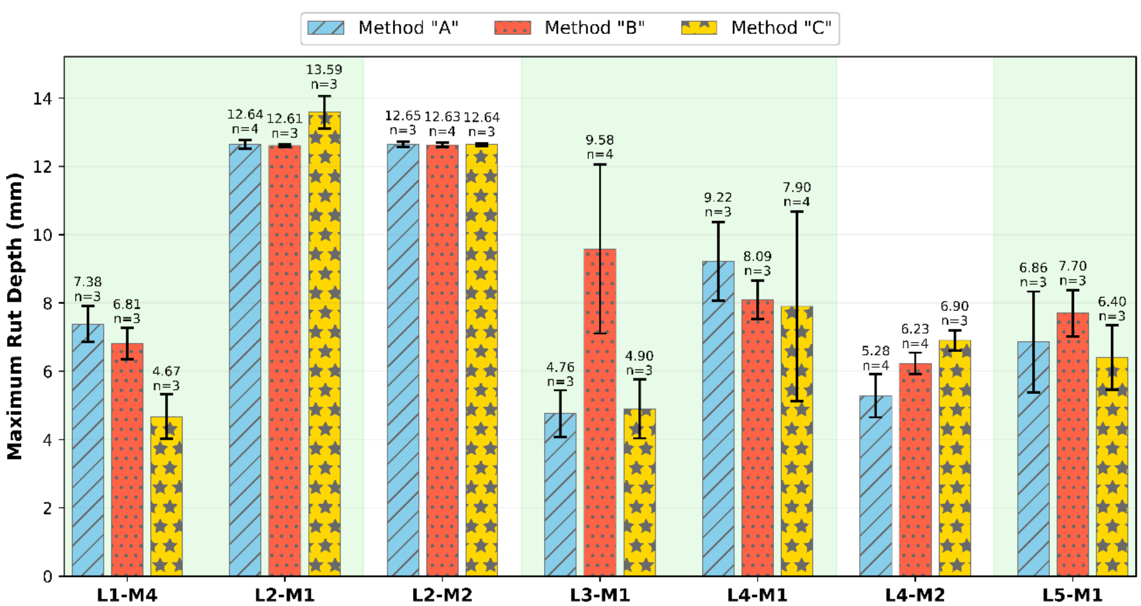


Figure A6.3. Summary of maximum rut depths for the evaluated mixtures (excluding outliers), categorized by specimen trimming method (source: FHWA).

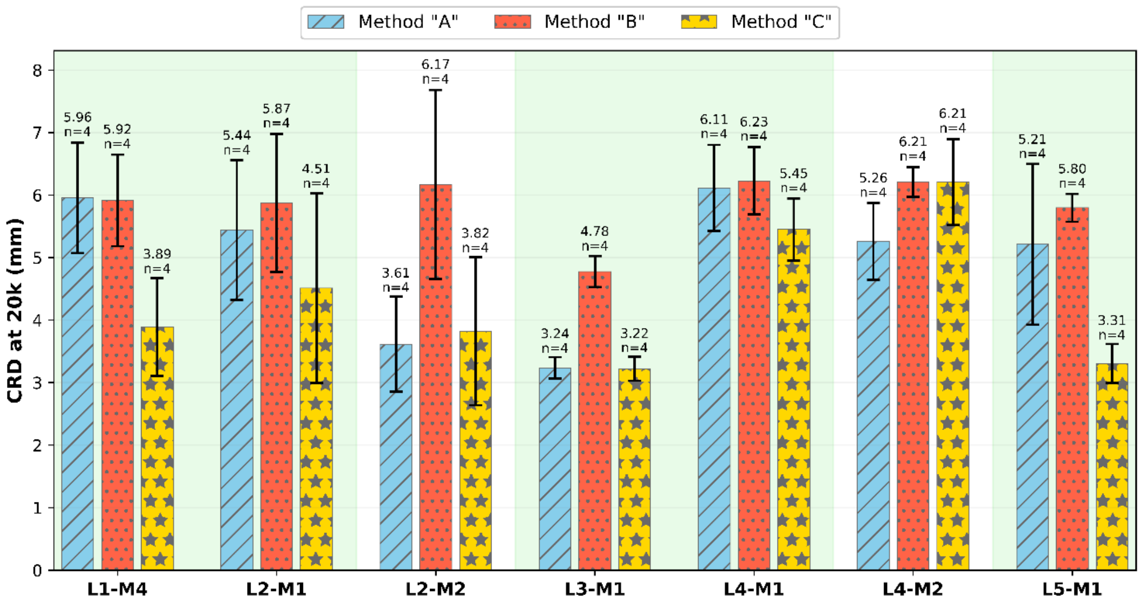


Figure A6.4. Summary of CRD20 for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

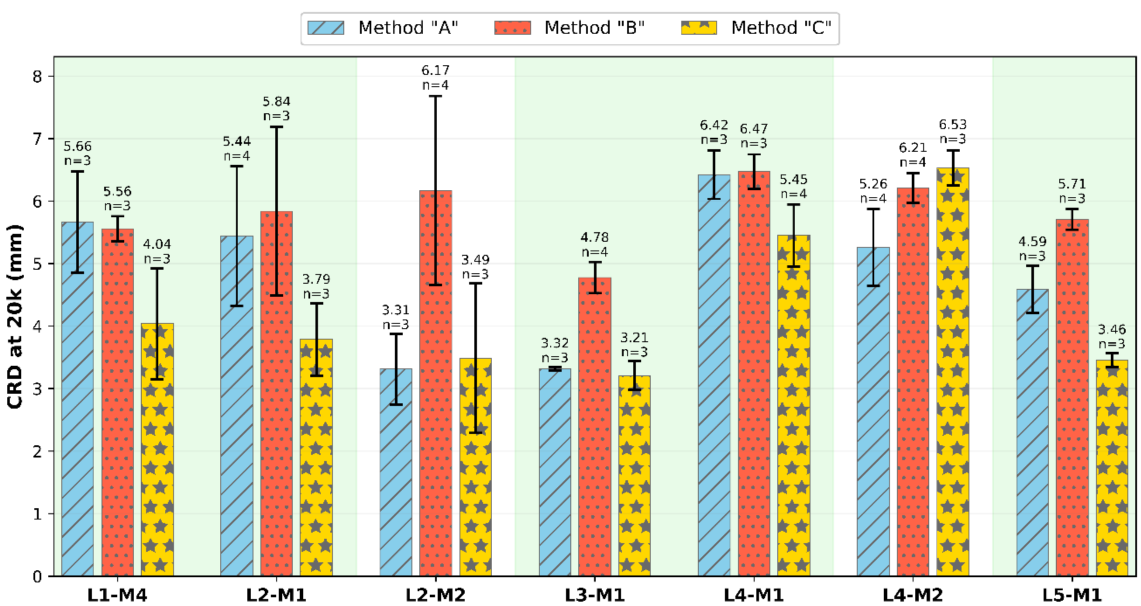


Figure A6.5. Summary of CRD20 for the evaluated mixtures (excluding outliers), categorized by specimen trimming method (source: FHWA).

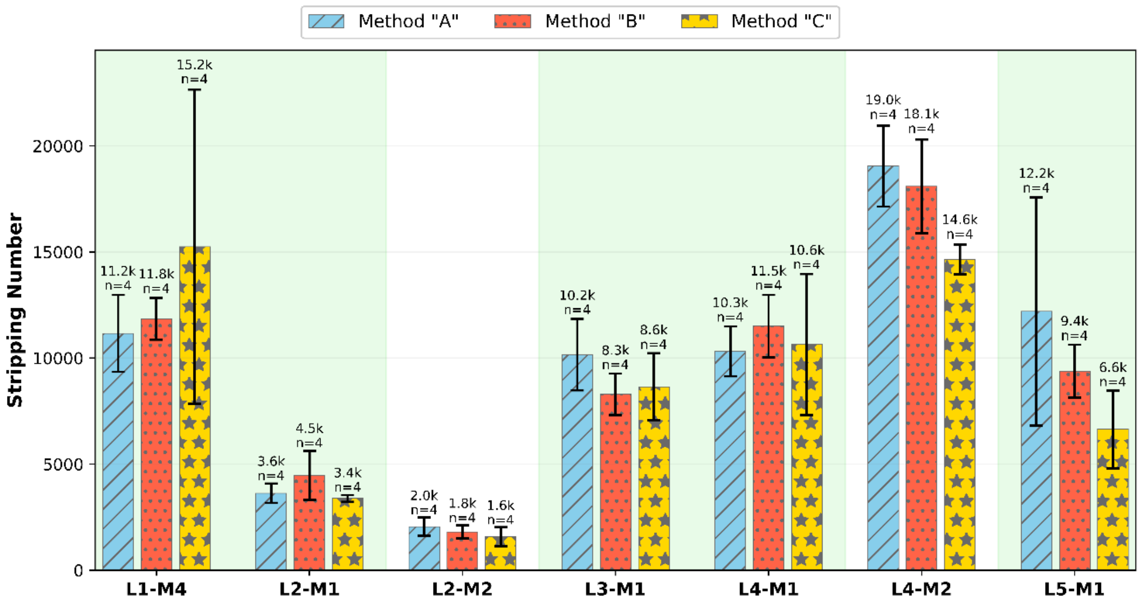


Figure A6.6. Summary of SN for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

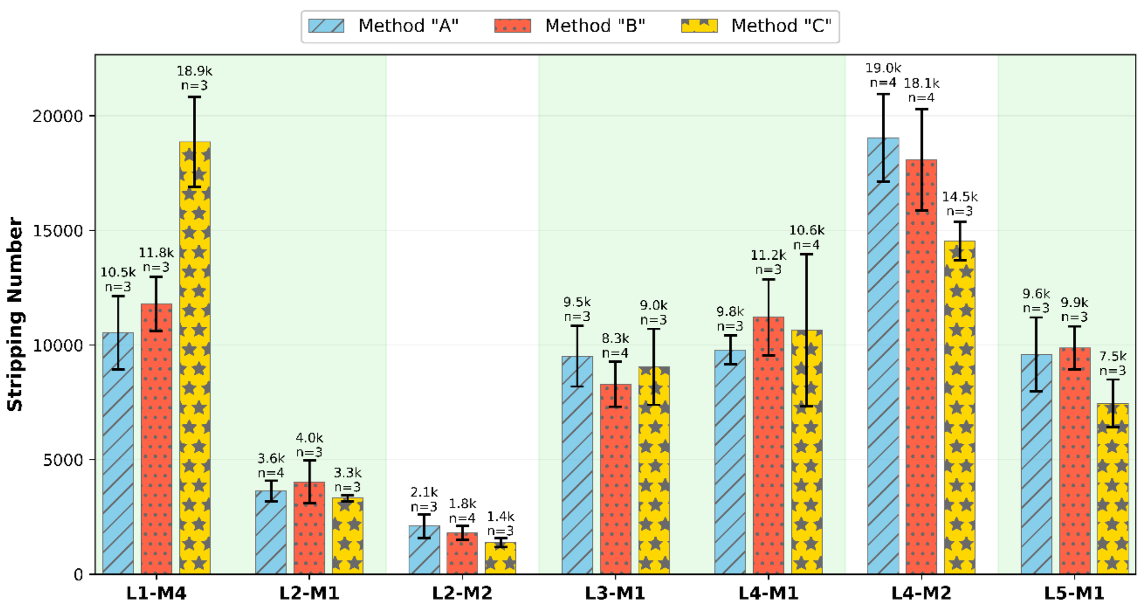


Figure A6.7. Summary of SN for the evaluated mixtures (excluding outliers), categorized by specimen trimming method (source: FHWA).

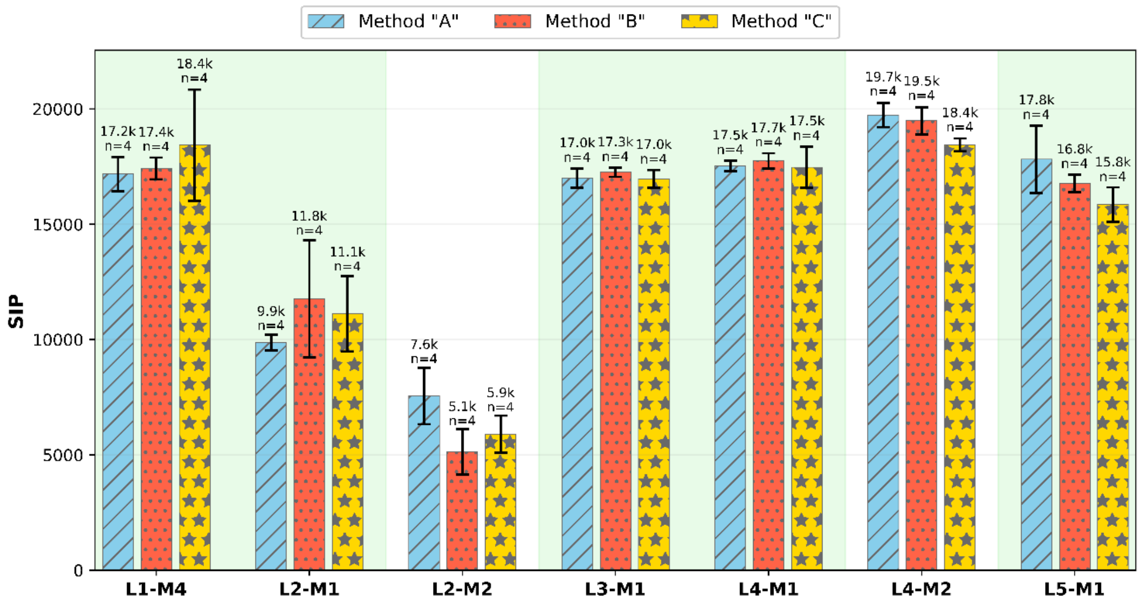


Figure A6.8. Summary of SIP for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

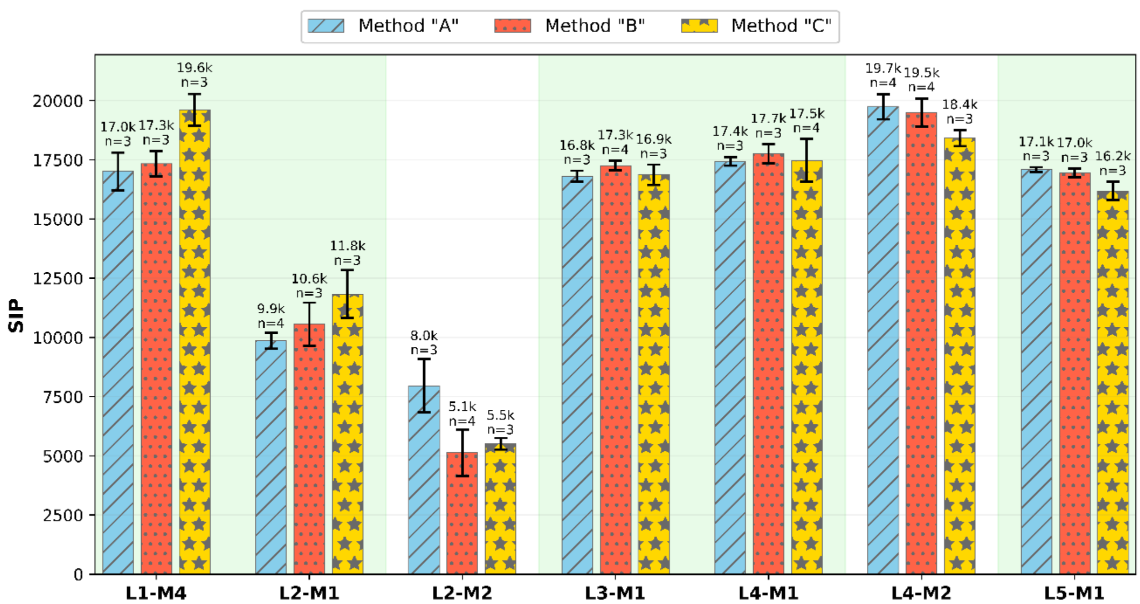


Figure A6.9. Summary of SIP for the evaluated mixtures (excluding outliers), categorized by specimen trimming method (source: FHWA).

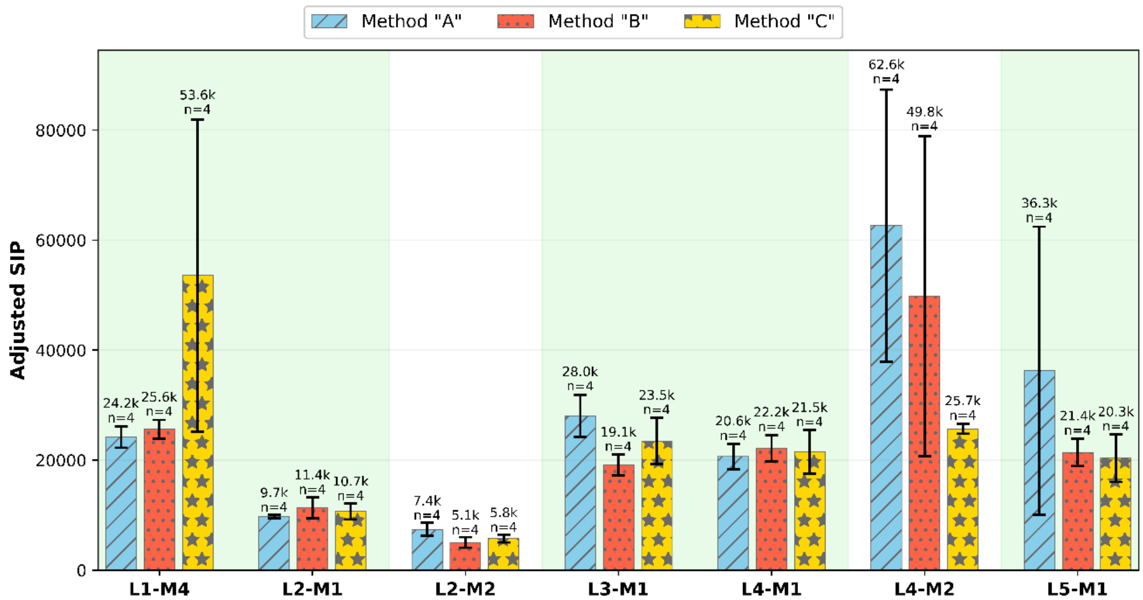


Figure A6.10. Summary of adjusted SIP for the evaluated mixtures based on all replicates and categorized by specimen trimming method (source: FHWA).

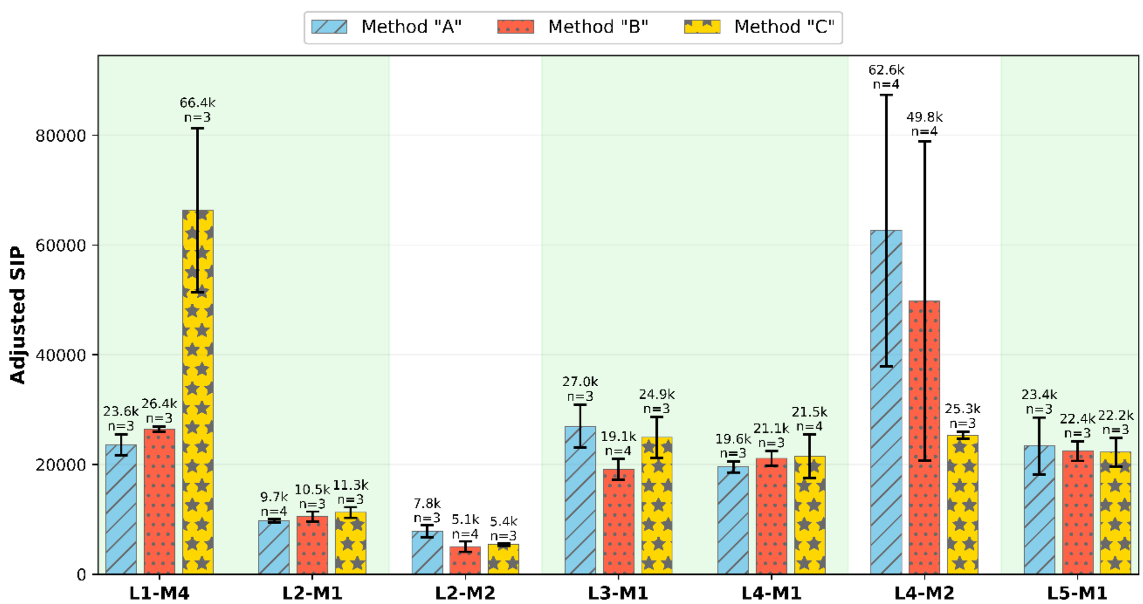


Figure A6.11. Summary of adjusted SIP for the evaluated mixtures (excluding outliers), categorized by specimen trimming method (source: FHWA).

A6.3 Equality Plots on the Effect of Trimming Method

This section includes the equality plots showing the effect of trimming method on different parameters calculated from 2PP modeling on HWTT rutting curves across evaluated mixtures. Each figure is structured as multi-panel grids. Within each grid, the performance metrics derived from specimens prepared using a trimming method are plotted against those utilizing another trimming method. Each datapoint on these equality plots indicates the average parameter value for the sample set subjected to certain trimming methods.

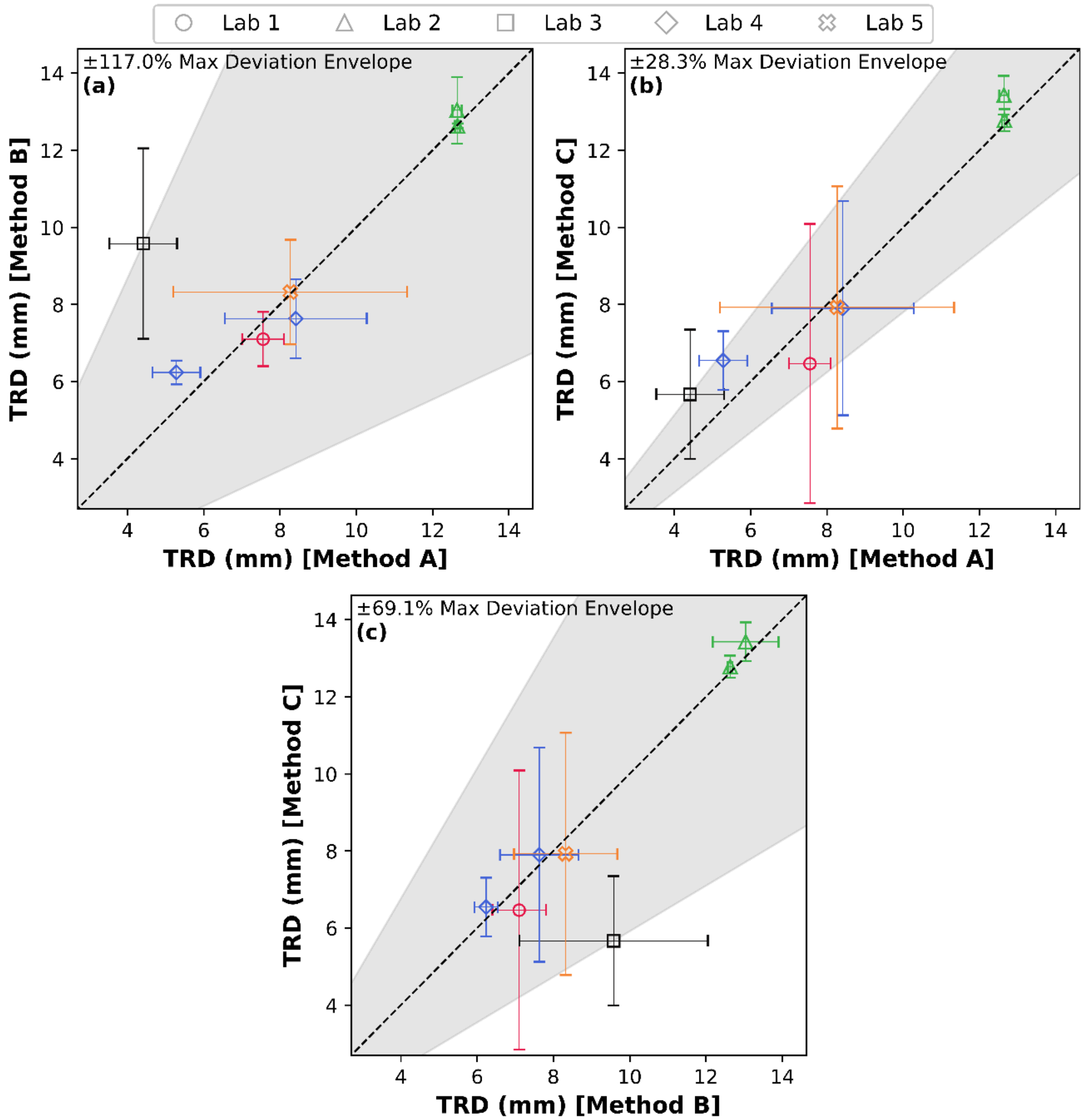


Figure A6.12. Equality plots comparing TRD of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

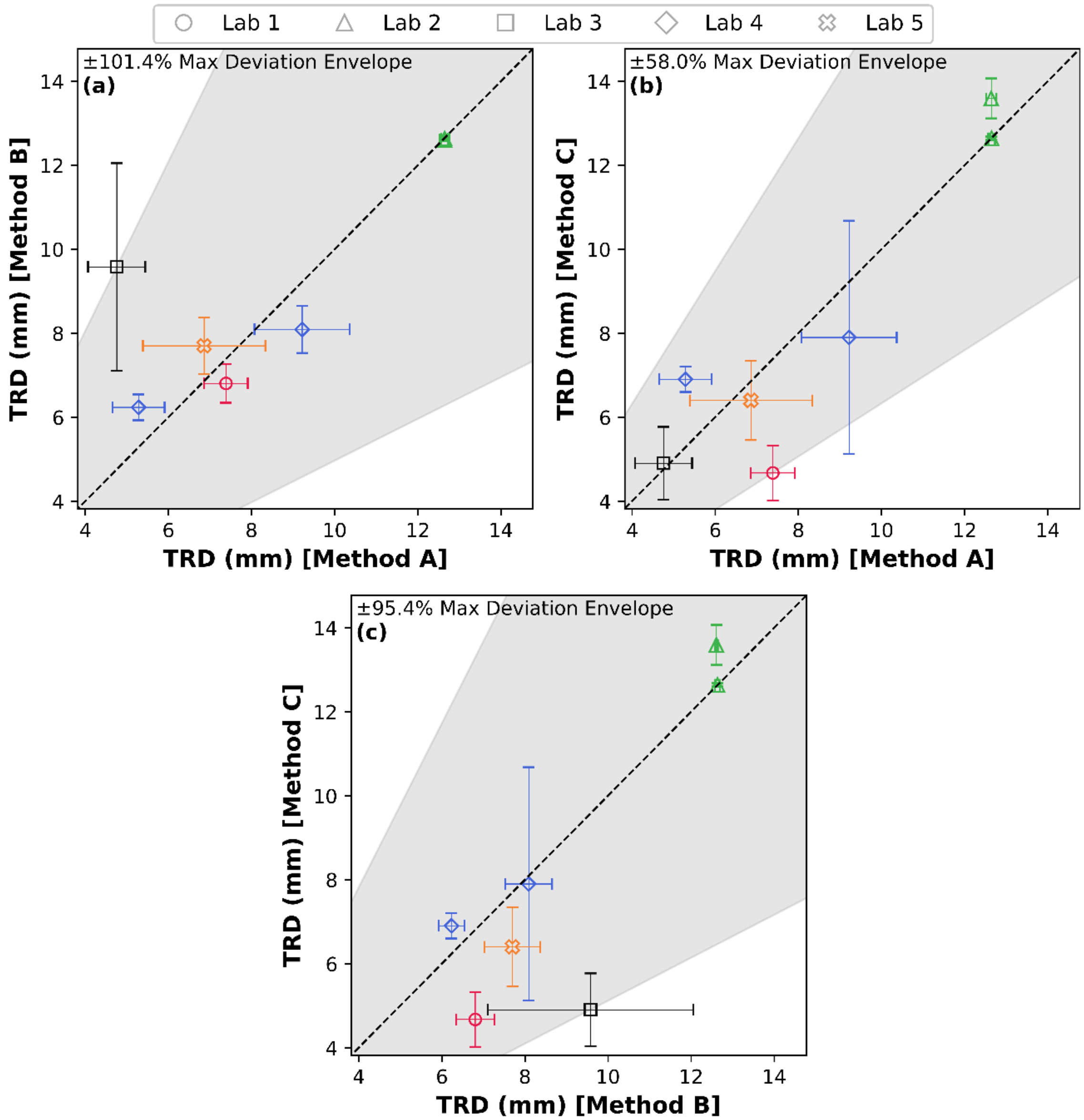


Figure A6.13. Equality plots comparing TRD of specimens prepared using different trimming methods (excluding outliers): (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

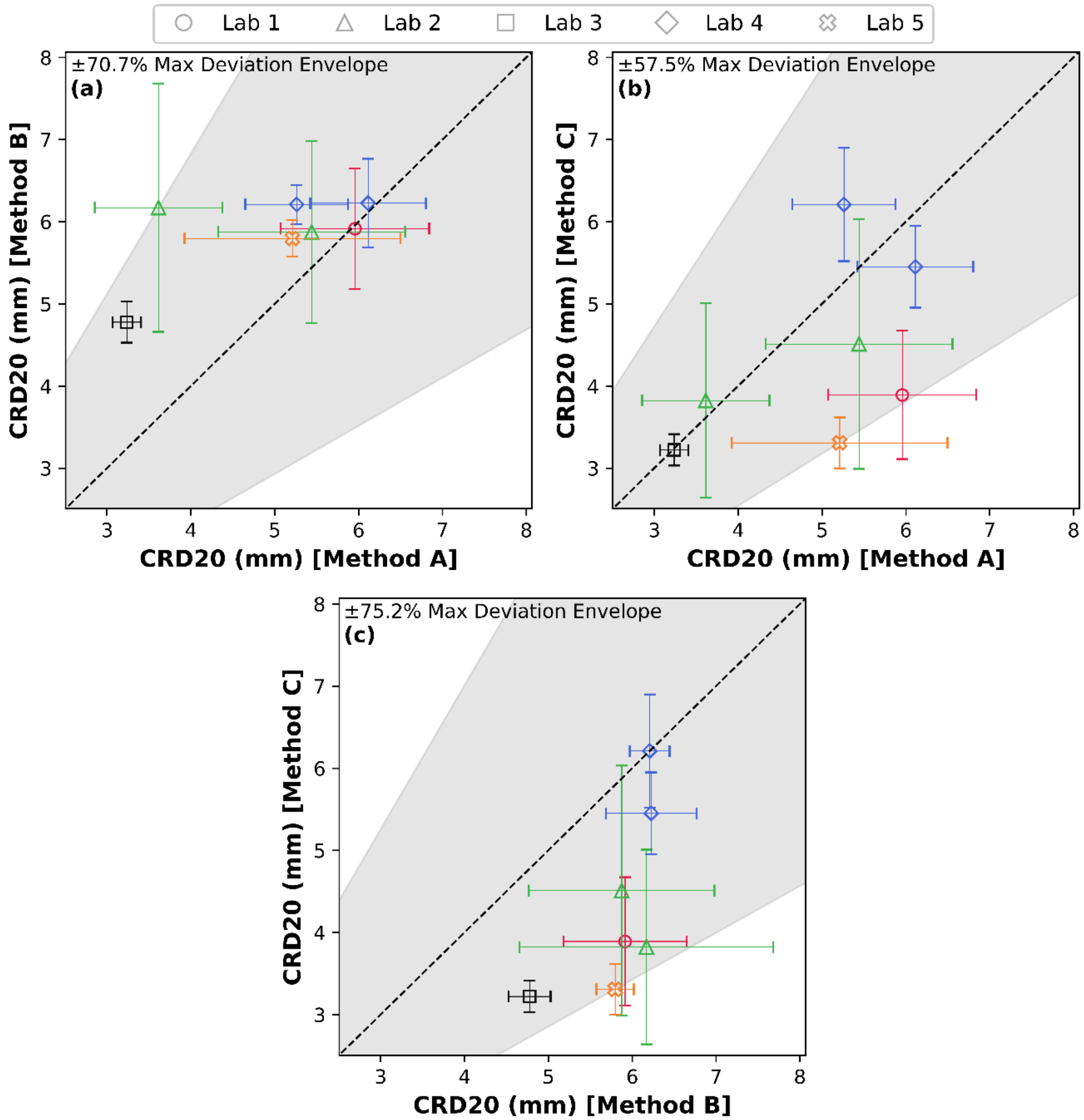


Figure A6.14. Equality plots comparing CRD20 of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

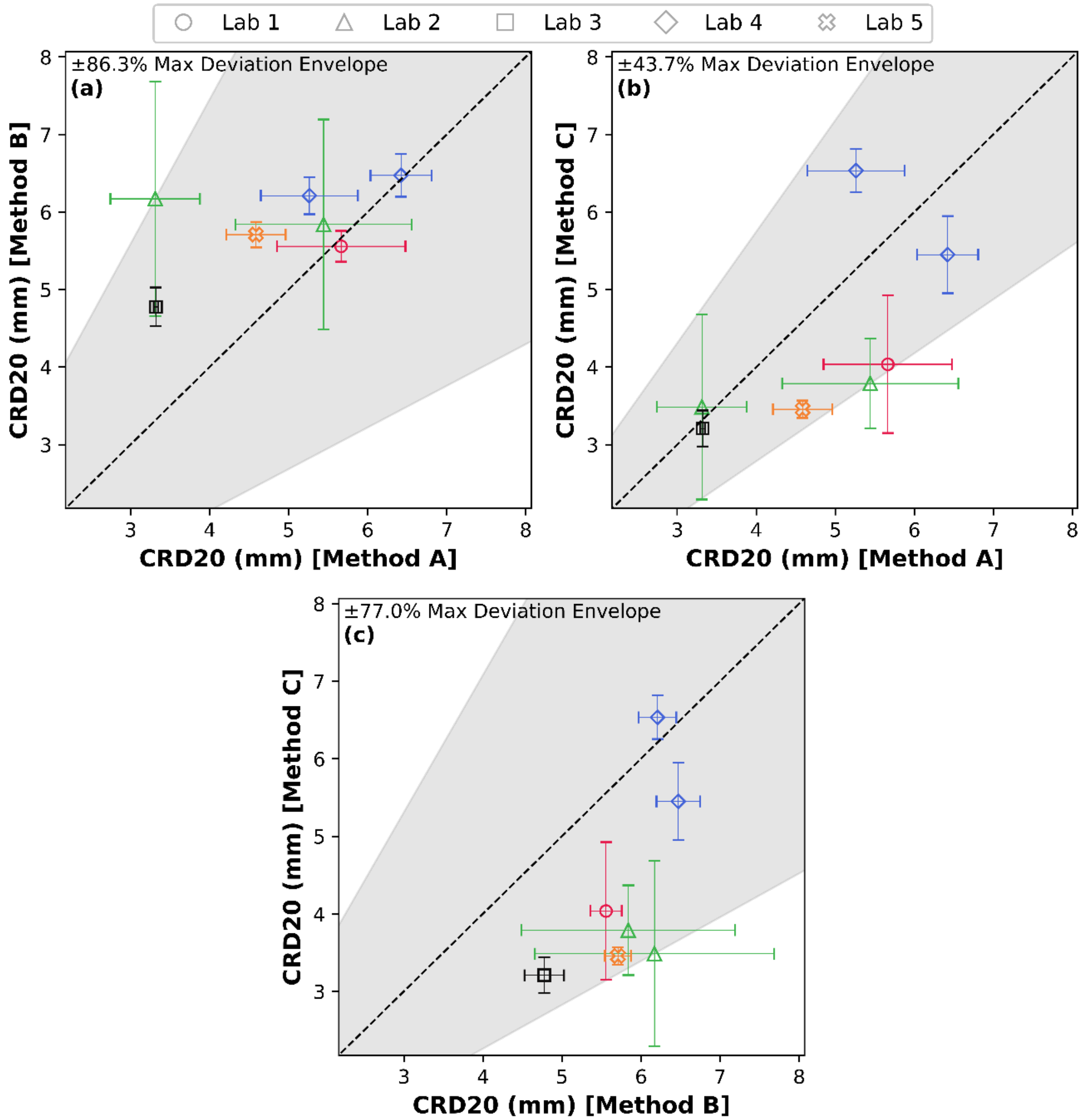


Figure A6.15. Equality plots comparing CRD20 of specimens prepared using different trimming methods (excluding outliers): (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

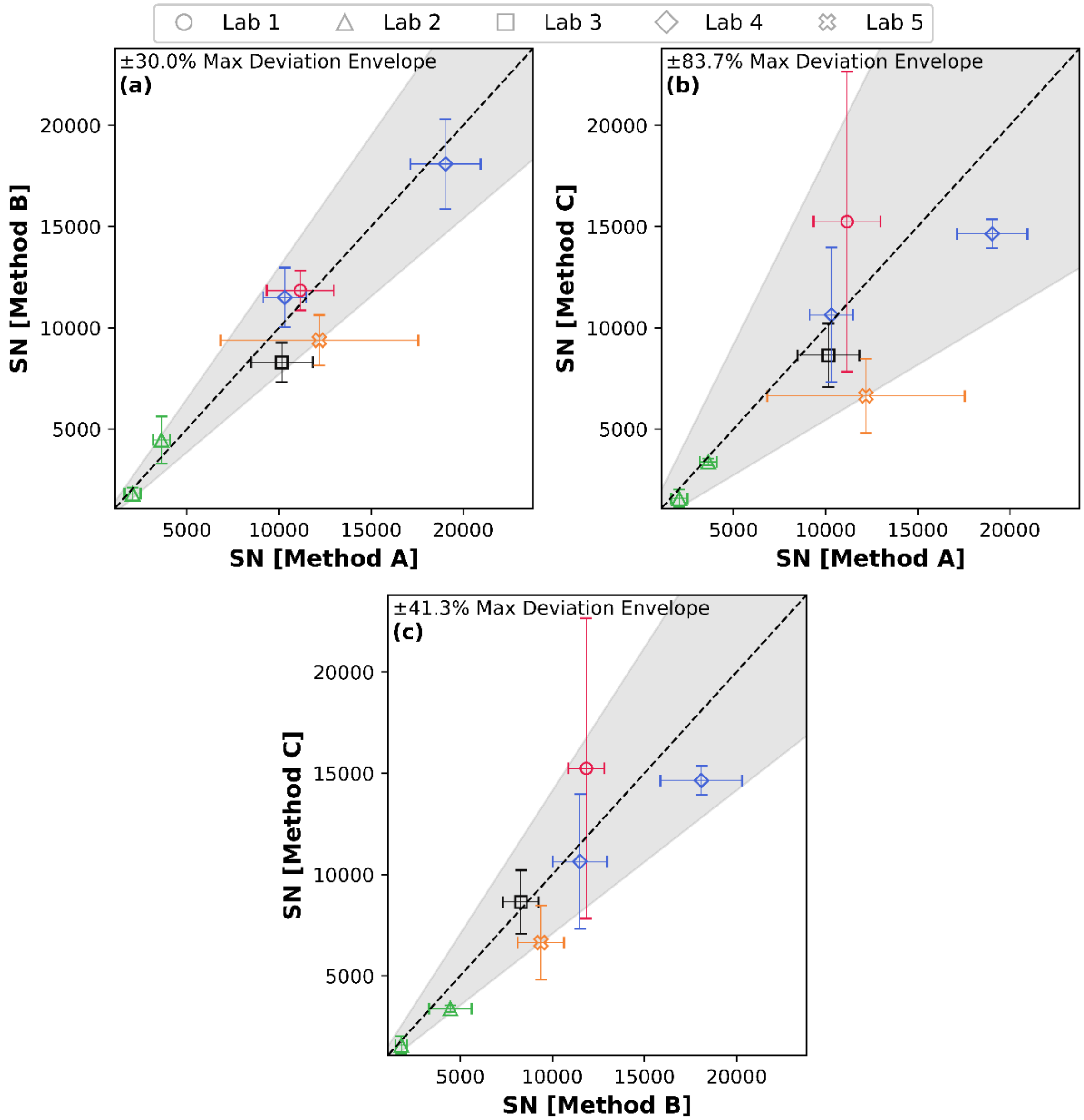


Figure A6.16. Equality plots comparing SN of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

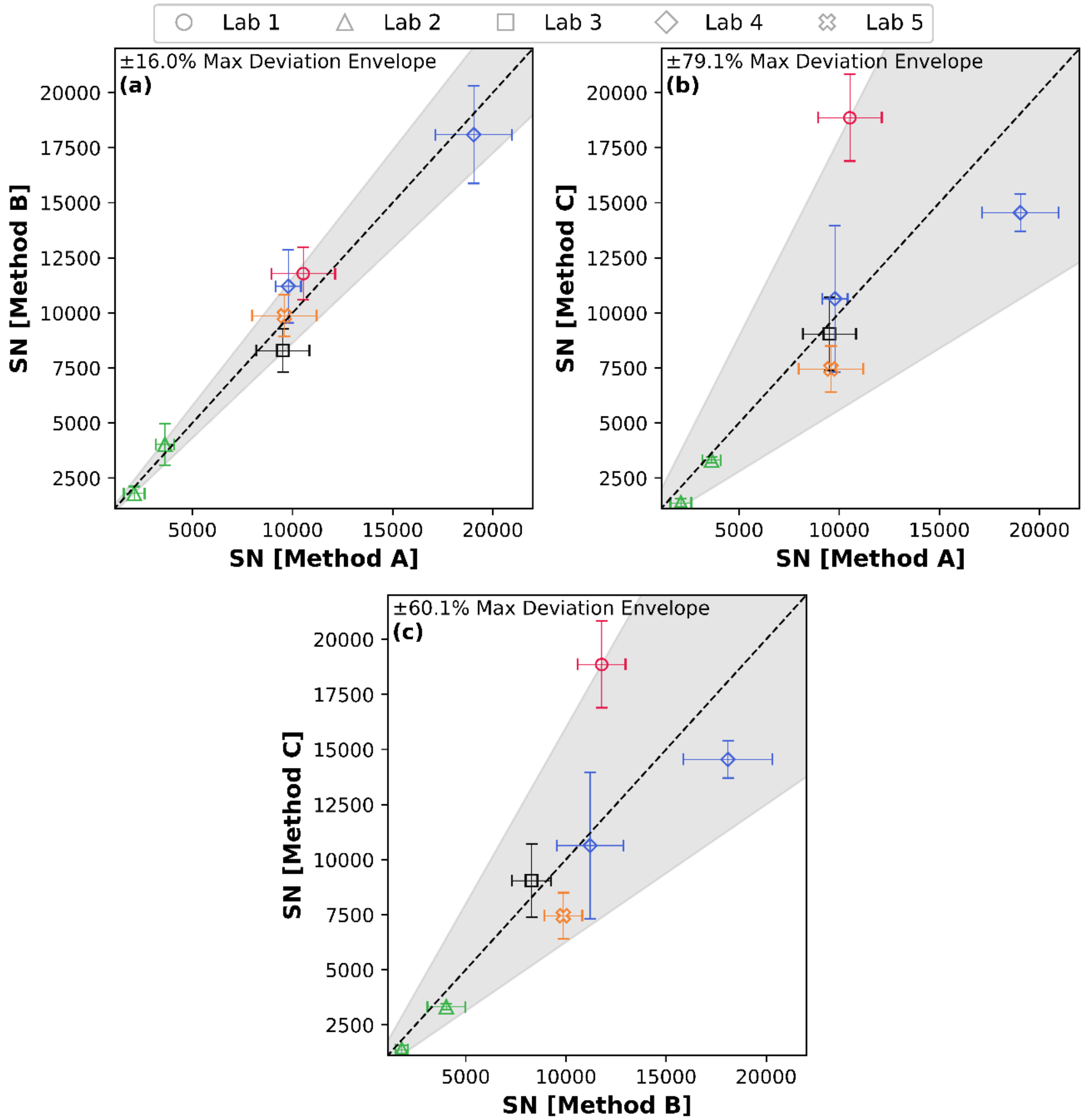


Figure A6.17. Equality plots comparing SN of specimens prepared using different trimming methods (excluding outliers): (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

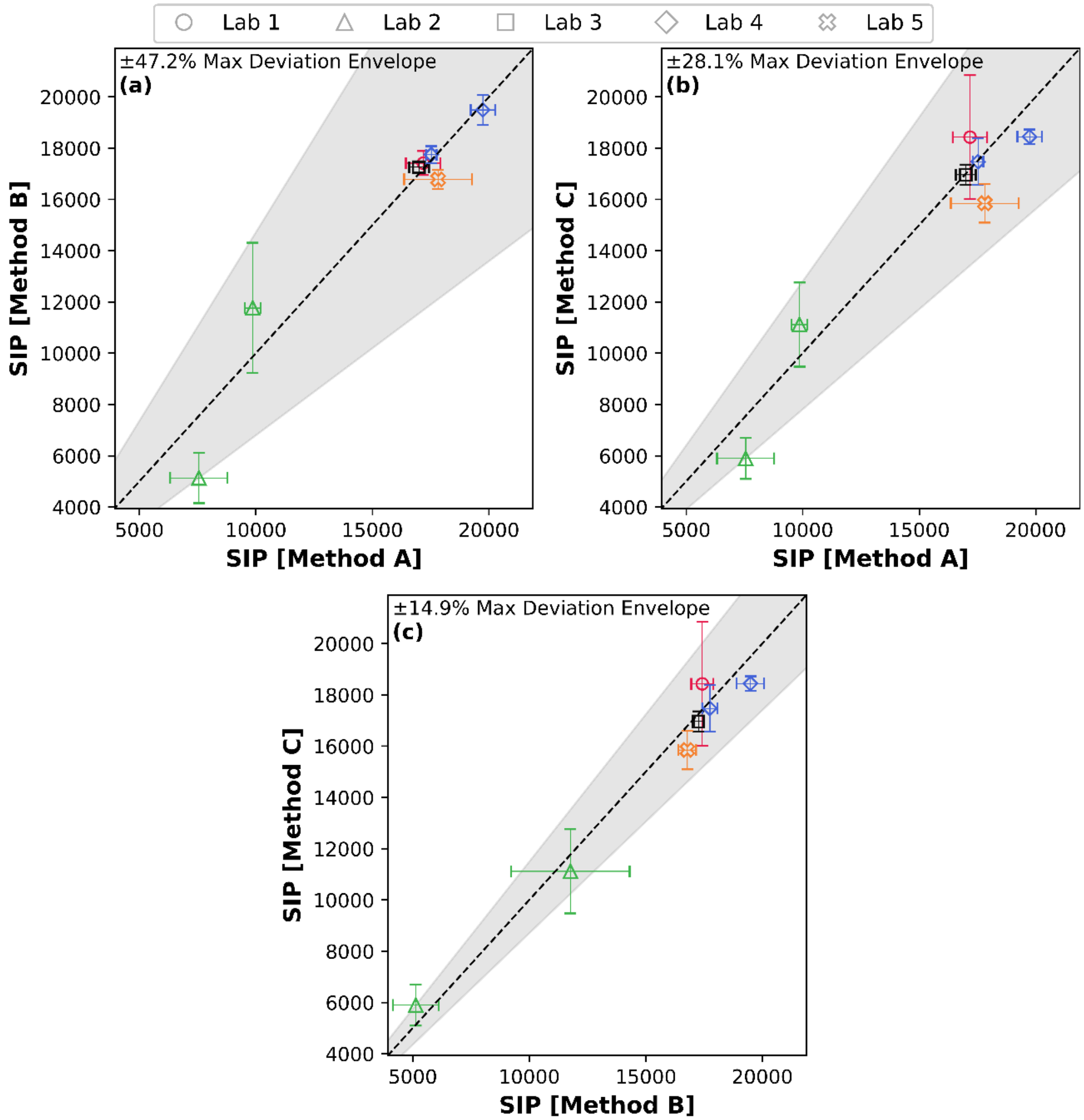


Figure A6.18. Equality plots comparing SIP of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

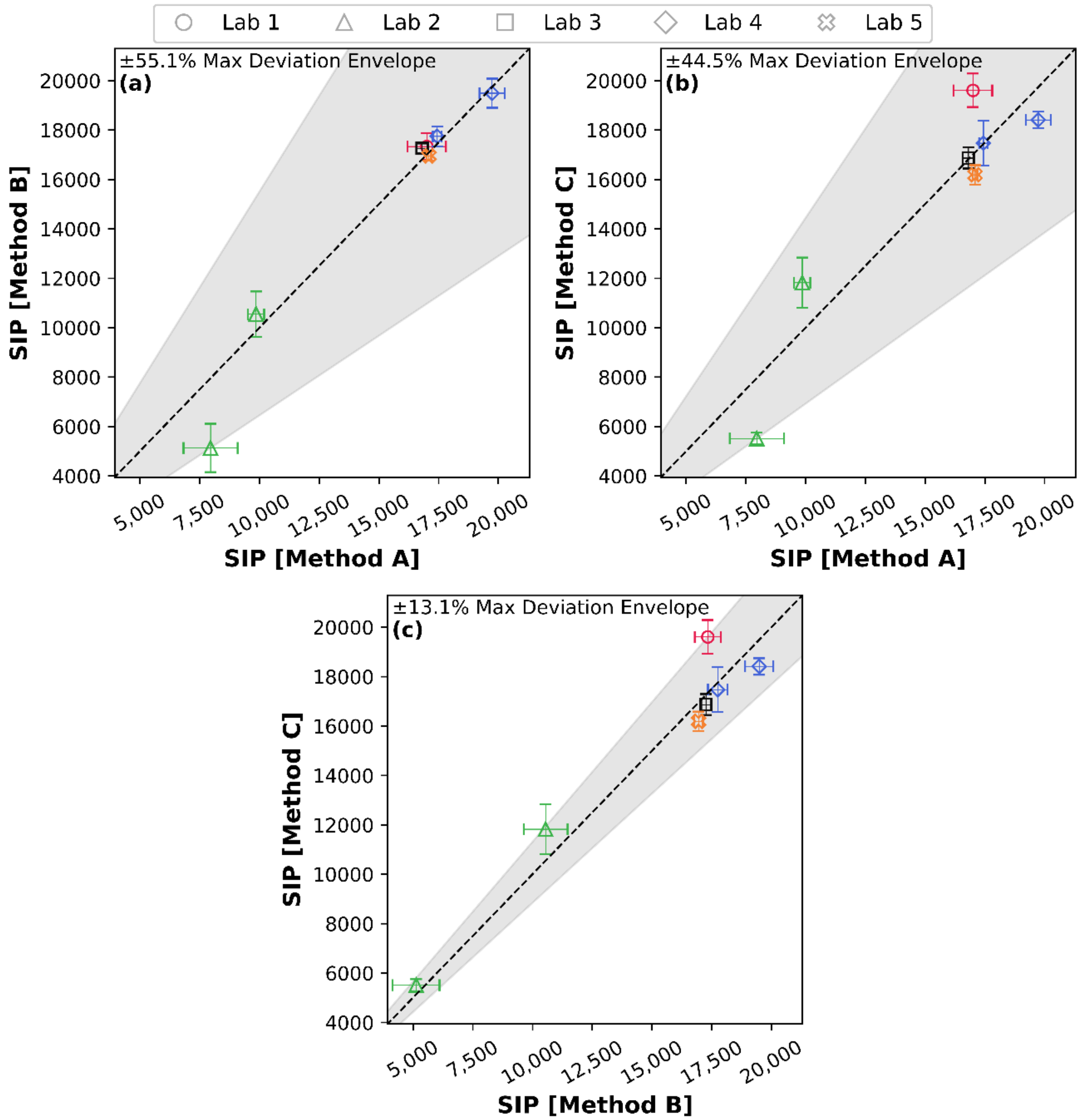


Figure A6.19. Equality plots comparing SIP of specimens prepared using different trimming methods (excluding outliers): (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

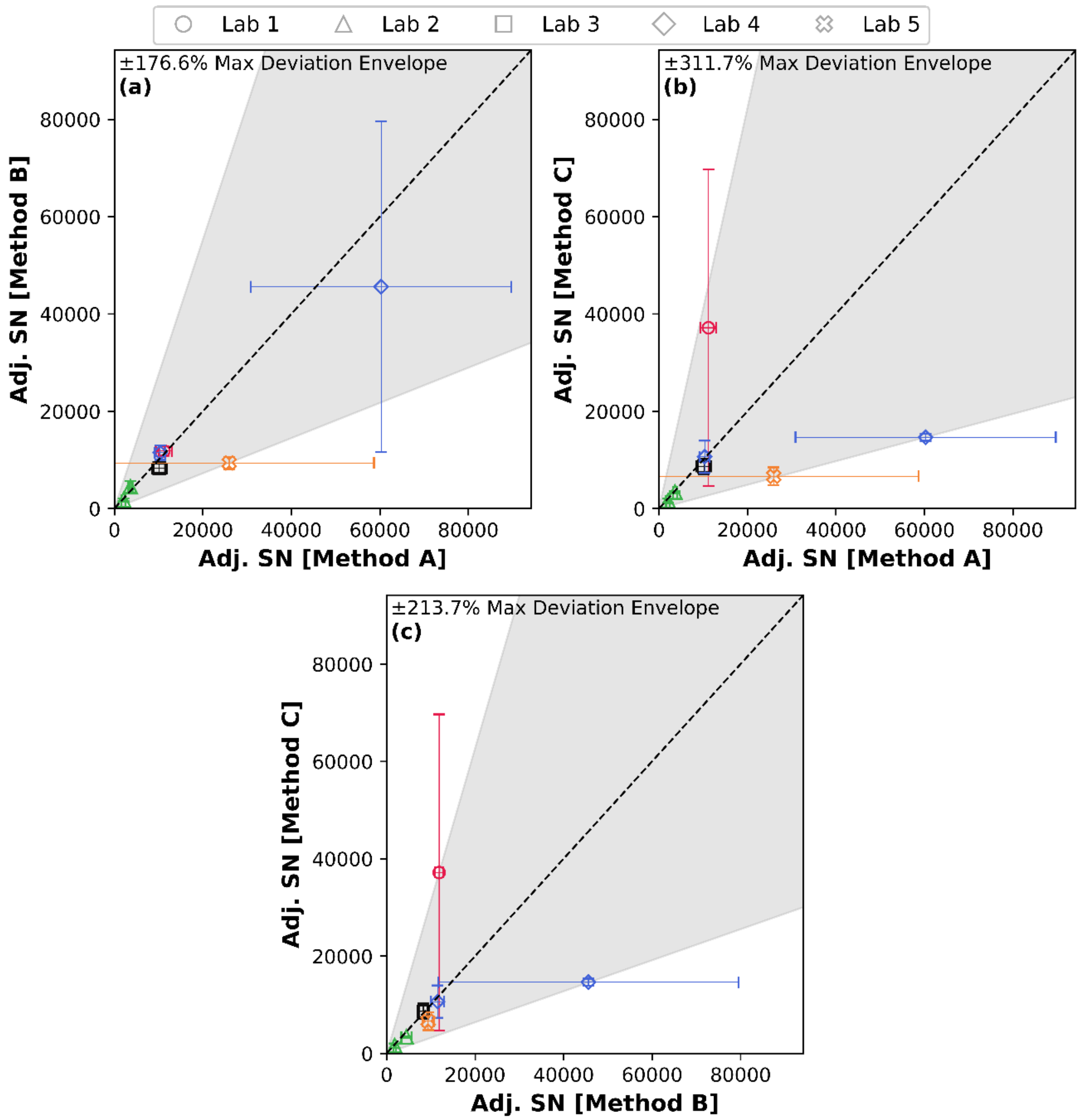


Figure A6.20. Equality plots comparing adjusted SIP of specimens prepared using different trimming methods based on all replicates: (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

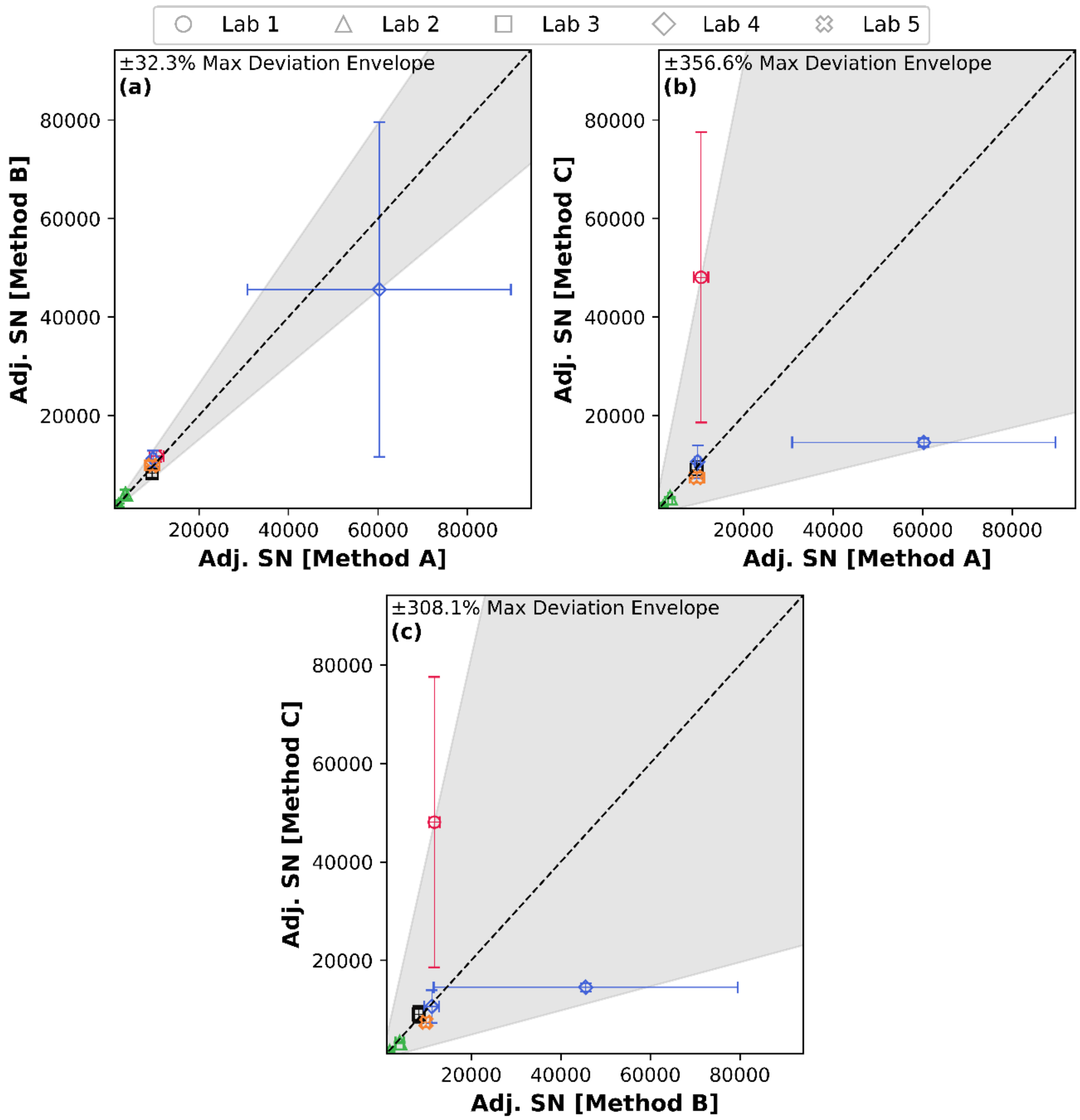


Figure A6.21. Equality plots comparing adjusted SIP of specimens prepared using different trimming methods (excluding outliers): (a) Method A vs. Method B, (b) Method A vs. Method C, (c) Method B vs. Method C (source: FHWA).

A6.4 One-Way ANOVA Analysis on the Effect of Trimming Method

To isolate and statistically assess the influence of the three-specimen trimming method (Methods A, B, and C) on an individual mixture basis, a one-way ANOVA was performed. Table A6.1 compiles the resulting F-statistics and p-values for the seven evaluated asphalt mixtures across the complete suite of HWTT performance metrics. It is important to emphasize the omnibus nature of the one-way ANOVA; a statistically significant outcome ($p \leq 0.05$) confirms only that the mean performance of at least two trimming procedures significantly differ from one another. The ANOVA inherently lacks the resolution to pinpoint exactly which specific trimming methods are driving this variance.

Table A6.1. Summary of one-way ANOVA results evaluating the effect of trimming method across all performance parameters.

Mixture ID	TRD		CRD20		SN		SIP		Adj. SIP	
	p-value	F-stat	p-value	F-stat	p-value	F-stat	p-value	F-stat	p-value	F-stat
Based on all replicates (including outliers)										
L1-M4	✗ 0.78	0.258	✓ 0.01	8.672	✗ 0.42	0.968	✗ 0.48	0.808	✗ 0.06	4.056
L2-M1	✗ 0.22	1.827	✗ 0.34	1.218	✗ 0.14	2.475	✗ 0.34	1.217	✗ 0.31	1.325
L2-M2	✗ 0.46	0.850	✓ 0.03	5.665	✗ 0.30	1.400	✓ 0.02	5.887	✓ 0.02	6.314
L3-M1	✓ 0.01	8.973	✓ 0.00	75.140	✗ 0.21	1.888	✗ 0.47	0.831	✓ 0.02	6.672
L4-M1	✗ 0.86	0.157	✗ 0.18	2.083	✗ 0.74	0.309	✗ 0.77	0.264	✗ 0.77	0.265
L4-M2	✓ 0.04	4.829	✗ 0.06	3.946	✓ 0.01	7.090	✓ 0.01	8.100	✗ 0.11	2.886
L5-M1	✗ 0.97	0.027	✓ 0.00	11.287	✗ 0.12	2.742	✗ 0.05	4.110	✗ 0.31	1.340
Based on filtered replicates (excluding outliers)										
L1-M4	✓ 0.00	20.092	✗ 0.05	5.013	✓ 0.00	23.173	✓ 0.01	12.898	✓ 0.00	22.555
L2-M1	✓ 0.00	13.615	✗ 0.11	3.154	✗ 0.38	1.107	✓ 0.03	5.730	✗ 0.09	3.507
L2-M2	✗ 0.95	0.055	✓ 0.03	6.241	✗ 0.11	3.084	✓ 0.01	9.508	✓ 0.01	9.644
L3-M1	✓ 0.01	9.130	✓ 0.00	65.048	✗ 0.49	0.793	✗ 0.15	2.549	✓ 0.03	6.070
L4-M1	✗ 0.66	0.435	✓ 0.02	7.004	✗ 0.77	0.274	✗ 0.81	0.219	✗ 0.65	0.454
L4-M2	✓ 0.00	11.294	✓ 0.01	8.730	✓ 0.03	5.480	✓ 0.02	6.295	✗ 0.17	2.199
L5-M1	✗ 0.39	1.114	✓ 0.00	63.125	✗ 0.10	3.435	✓ 0.01	10.744	✗ 0.92	0.088

A comprehensive review of Table A6.1 reveals several distinct behavioral trends regarding parameter sensitivity to the trimming methods. First, while practitioners often hesitate to discard any data points due to the limited number of experimental replicates (typically four per mixture) and the resources needed to run additional HWTT specimens, the application of the outlier detection algorithm—which is strictly capped to remove a maximum of one anomalous replicate to preserve viable sample sizes—demonstrated a profound impact on the statistical outcomes. Across all evaluated parameters, excluding these outliers consistently resulted in a higher number of mixtures exhibiting a significant trimming effect. For example, when analyzing the dataset inclusive of all replicates, the conventional SIP parameter indicated that the trimming method was significant for only two mixtures (L2-M2 and L4-M2). However, once the statistical outliers were filtered, five out of the seven mixtures demonstrated a statistically significant trimming effect on the SIP. This filtered outcome presents a more logically consistent narrative that aligns closely with the behaviors of the other stripping-related metrics: SN and Adjusted SIP.

A second critical observation from Table A6.1 refers to the varying degrees of mixture-specific sensitivity to the trimming methods. Based on the filtered dataset (excluding outliers), mixtures L4-M2 and L1-M4 exhibited the highest overall sensitivity to the trimming methods. Specifically for mixture L4-M2, every evaluated performance parameter, except adjusted SIP, was significantly impacted by the trimming method. Conversely, mixture L4-M1 exhibited the exact opposite behavioral trend; within the filtered dataset, the CRD20 was the only parameter statistically impacted by the trimming method, while the remaining metrics showed no significant variance. This stark contrast in responsiveness confirms that the effect of specimen trimming is a mixture-dependent phenomenon.

A6.5. Post-Hoc Analysis on Effect of Trimming Method

As noted previously, the omnibus one-way ANOVA solely indicates the presence of a statistically significant variance caused by the trimming method, without identifying the specific contrasting pairs. To explicitly determine which specific trimming methods produced significantly different performance metrics from one another for a given mixture, a Games-Howell post-hoc analysis was subsequently performed. The detailed results of these pairwise comparisons are presented in Table A6.2 through Table A6.8. Within these tables, all possible combinations of the trimming procedures (Methods A, B, and C) are cross-compared, and their corresponding p-values are reported for both the complete dataset (inclusive of all experimental replicates) and the filtered dataset (exclusive of statistical outliers). Any pairwise comparison yielding a p-value of 0.05 or less confirms a statistically significant difference between the performance means of those two specific methods. To facilitate rapid visual interpretation, consistent with the reporting format established in Appendix A5.5, an "Is Diff?" column is included in each table, utilizing a checkmark (✓) to denote a statistically significant difference and a cross symbol (✗) to indicate a lack of statistical significance.

Table A6.2. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L1-M4.

Group 1	**Group 2**	**TRD**		**CRD20**		**SN**		**SIP**		**Adjusted SIP**	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
Based on all replicates (including outliers)											
A	B	0.60	✗	1.00	✗	0.79	✗	0.85	✗	0.56	✗
A	C	0.83	✗	0.03	✓	0.59	✗	0.62	✗	0.53	✗
B	C	0.94	✗	0.02	✓	0.67	✗	0.71	✗	0.54	✗
Based on filtered replicates (excluding outliers)											
A	B	0.41	✗	0.97	✗	0.57	✗	0.83	✗	0.21	✗
A	C	0.01	✓	0.16	✗	0.01	✓	0.03	✓	0.53	✗
B	C	0.03	✓	0.16	✗	0.02	✓	0.03	✓	0.54	✗

Table A6.3. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L2-M1.

Group 1	**Group 2**	**TRD**		**CRD20**		**SN**		**SIP**		**Adjusted SIP**	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
Based on all replicates (including outliers)											
A	B	0.67	✗	0.85	✗	0.44	✗	0.41	✗	0.35	✗
A	C	0.10	✗	0.61	✗	0.61	✗	0.40	✗	0.52	✗
B	C	0.73	✗	0.38	✗	0.29	✗	0.91	✗	0.83	✗
Based on filtered replicates (excluding outliers)											
A	B	0.85	✗	0.91	✗	0.79	✗	0.52	✗	0.51	✗
A	C	0.12	✗	0.12	✗	0.49	✗	0.13	✗	0.19	✗
B	C	0.12	✗	0.19	✗	0.52	✗	0.34	✗	0.61	✗

Table A6.4. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L2-M2.

Group 1	Group 2	TRD		CRD20		SN		SIP		Adjusted SIP	
		p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*
Based on all replicates (including outliers)											
A	B	0.89	✗	0.07	✗	0.65	✗	0.05	✗	0.05	✓
A	C	0.69	✗	0.95	✗	0.34	✗	0.15	✗	0.13	✗
B	C	0.62	✗	0.11	✗	0.69	✗	0.49	✗	0.51	✗
Based on filtered replicates (excluding outliers)											
A	B	0.96	✗	0.05	✗	0.70	✗	0.05	✗	0.05	✗
A	C	0.98	✗	0.97	✗	0.23	✗	0.11	✗	0.11	✗
B	C	0.99	✗	0.10	✗	0.15	✗	0.76	✗	0.77	✗

Table A6.5. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L3-M1.

Group 1	Group 2	TRD		CRD20		SN		SIP		Adjusted SIP	
		p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*	p-value	Is Diff?*
Based on all replicates (including outliers)											
A	B	0.04	✓	0.00	✓	0.23	✗	0.56	✗	0.03	✓
A	C	0.45	✗	0.99	✗	0.44	✗	0.99	✗	0.31	✗
B	C	0.10	✗	0.00	✓	0.92	✗	0.44	✗	0.25	✗
Based on filtered replicates (excluding outliers)											
A	B	0.05	✗	0.00	✓	0.45	✗	0.11	✗	0.11	✗
A	C	0.97	✗	0.72	✗	0.92	✗	0.98	✗	0.80	✗
B	C	0.05	✗	0.00	✓	0.78	✗	0.43	✗	0.18	✗

Table A6.6. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L4-M1.

Group 1	Group 2	TRD		CRD20		SN		SIP		Adjusted SIP	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
Based on all replicates (including outliers)											
A	B	0.75	✗	0.96	✗	0.47	✗	0.55	✗	0.65	✗
A	C	0.95	✗	0.34	✗	0.98	✗	0.99	✗	0.93	✗
B	C	0.98	✗	0.17	✗	0.89	✗	0.84	✗	0.95	✗
Based on filtered replicates (excluding outliers)											
A	B	0.40	✗	0.98	✗	0.46	✗	0.53	✗	0.34	✗
A	C	0.69	✗	0.07	✗	0.88	✗	1.00	✗	0.65	✗
B	C	0.99	✗	0.04	✓	0.95	✗	0.85	✗	0.98	✗

Table A6.7. Games-Howell pairwise comparison of performance metrics across trimming methods for Mixture L4-M2.

Group 1	Group 2	TRD		CRD20		SN		SIP		Adjusted SIP	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
Based on all replicates (including outliers)											
A	B	0.10	✗	0.10	✗	0.80	✗	0.81	✗	0.42	✗
A	C	0.10	✗	0.18	✗	0.03	✓	0.02	✓	0.19	✗
B	C	0.75	✗	1.00	✗	0.10	✗	0.06	✗	0.34	✗
Based on filtered replicates (excluding outliers)											
A	B	0.10	✗	0.10	✗	0.80	✗	0.81	✗	0.42	✗
A	C	0.02	✓	0.04	✓	0.03	✓	0.02	✓	0.19	✗
B	C	0.08	✗	0.34	✗	0.09	✗	0.06	✗	0.34	✗

Table A6.8. Games-Howell pairwise comparison of performance metrics across oven heating protocols for Mixture L5-M1.

Group 1	Group 2	TRD		CRD20		SN		SIP		Adjusted SIP	
		p-value	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***	**p-value**	**Is Diff?***
Based on all replicates (including outliers)											
A	B	1.00	✗	0.68	✗	0.61	✗	0.44	✗	0.57	✗
A	C	0.99	✗	0.11	✗	0.25	✗	0.14	✗	0.54	✗
B	C	0.97	✗	0.00	✓	0.12	✗	0.17	✗	0.91	✗
Based on filtered replicates (excluding outliers)											
A	B	0.68	✗	0.04	✓	0.96	✗	0.55	✗	0.95	✗
A	C	0.90	✗	0.05	✗	0.26	✗	0.09	✗	0.94	✗
B	C	0.25	✗	0.00	✓	0.09	✗	0.11	✗	0.99	✗

A6.6 Two-Way ANOVA Results on Effect of Trimming Method

This section includes the detailed tabulated results of two-way ANOVA analysis based on different parameters, with and without including the outliers.

Table A5.9. Results of two-way ANOVA based on CRD20 parameter for Task 4 using filtered replicates (excluding outliers).

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	144.92	2	140.834	0.00
Mixture	213.07	6	69.022	0.00
Combined Effect	60.72	12	9.835	0.00
Residual	24.70	48		

Table A5.10. Results of two-way ANOVA based on TRD parameter for Task 4 using all replicates.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	13.81	2	2.452	0.09
Mixture	7,515.57	6	444.737	0.00
Combined Effect	65.81	12	1.947	0.04
Residual	177.44	63		

Table A5.11. Results of two-way ANOVA based on TRD parameter for Task 4 using filtered replicates (excluding outliers).

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	0.45	2	0.185	0.83
Mixture	4,711.76	6	647.556	0.00
Combined Effect	74.08	12	5.091	0.00
Residual	58.21	48		

Table A5.12. Results of two-way ANOVA based on SN parameter for Task 4 using all replicates.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	4.70e+7	2	3.954	0.02
Mixture	1.13e+10	6	317.29	0.00
Combined Effect	1.59e+8	12	2.22	0.02
Residual	3.75e+8	63		

Table A5.13. Results of two-way ANOVA based on SN parameter for Task 4 using filtered replicates (excluding outliers).

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	2.03e+7	2	4.565	0.02
Mixture	3.65e+9	6	274.001	0.00
Combined Effect	1.08e+8	12	4.062	0.00
Residual	1.07e+8	48		

Table A5.14. Results of two-way ANOVA based on SIP parameter for Task 4 using all replicates.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	1.46e+7	2	6.412	0.00
Mixture	2.99e+9	6	438.417	0.00
Combined Effect	2.89e+7	12	2.119	0.03
Residual	7.16e+7	63		

Table A5.15. Results of two-way ANOVA based on SIP parameter for Task 4 using filtered replicates (excluding outliers).

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	2.38e+6	2	3.111	0.05
Mixture	1.93e+9	6	839.415	0.00
Combined Effect	2.28e+8	12	4.957	0.00
Residual	1.84e+7	48		

Table A5.16. Results of two-way ANOVA based on adjusted SIP parameter for Task 4 using all replicates.

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	3.98e+9	2	0.488	0.06
Mixture	7.02e+12	6	286.786	0.00
Combined Effect	1.37e+11	12	2.799	0.00
Residual	2.57e+11	63		

Table A5.17. Results of two-way ANOVA based on adjusted SIP parameter for Task 4 using filtered replicates (excluding outliers).

Parameter	SSE	Degree of Freedom	F-statistic	p-value
Conditioning Method	3.38e+9	2	0.374	0.07
Mixture	1.59e+13	6	586.025	0.00
Combined Effect	1.72e+11	12	3.176	0.00
Residual	2.17e+11	48		