Conférence Européenne des Directeurs des Routes

Conference of European Directors of Roads

DRaT – Development of the Ravelling Test

Report on Analysis of Results

Deliverable 8

July 2017

Project Coordinator: TRL Limited, United Kingdom

Project Partners: Netherlands Organisation for Applied Sci. Research, The Netherlands

Belgian Road Research Centre, Belgium

BAM Infra Asfalt, The Netherlands

Heijmans Integrale Projecten, The Netherlands

IFSTTAR, France

Technische Universität Darmstadt, Germany

ISAC - RWTH Aachen University, Germany

Call 2014: Asset Management and Maintenance

Development of the Ravelling Test

Deliverable 8 – Report on Analysis of Results

Due date of deliverable: 30/06/2017

Actual submission date: 12/07/2017

Start date of project: 01/09/2015 End date of project: 31/08/2017

Authors of this deliverable:

Eric Schoen - Netherlands Organisation for Applied Sci. Research, The Netherlands

Dave van Vliet - Netherlands Organisation for Applied Sci. Research, The Netherlands

Steven Mookhoek - Netherlands Organisation for Applied Sci. Research, The Netherlands

Nadieh Meinen - Netherlands Organisation for Applied Sci. Research, The Netherlands

Project coordinator: Arthur Hannah, TRL

PEB Project Managers: Arash Khojinian (HE) and Franz Bommert (BASt)

Version: final

Executive summary

Recently, test methods for four scuffing devices have been written up as a draft technical specification by Comité Européen de Normalisation (CEN) as prCEN/TS 12697-50, Resistance to scuffing. The CEDR-DRaT project evaluates the test methods and the results produced for the four scuffing devices to make out whether the methods can lead to a harmonised test method of which the technical specification can be converted into a test standard.

This report describes the work done in Work Package 4 'Analysis' and contains the statistical analysis of the combined results obtained within the comparative study within the DRaT project. (De Visscher, 2017). The objectives of the Work Package were the following:

- Determine the precision of the results as the repeatability within each scuffing device and the similarity (reproducibility) of the results from similar scuffing devices.
- Determine the potential to correlate or unify the results of the different scuffing devices to a common measure.
- Determine the potential of the different scuffing devices to discriminate between asphalt qualities.

The collected individual results from each device were used to determine the variation of similar samples under near-homogeneous conditions (repeatability) of the test for the different asphalt types. Statistical techniques were employed to find potential outliers. In addition, scaling factors were calculated to convert the outcome of one device to the outcome of another device, and it was checked whether these scaling factors and the damage evolution in time per device depend on the asphalt mixture. Using the data collected for the different asphalt materials, with variation in standard ("good") and low-temperature-compaction/low-bitumen-designed mixtures ("bad"), the power to detect significant differences between the test results of the scuffing devices was established.

The following conclusions were drawn:

- The scuffing devices or methods cannot be used interchangeably because the devices' discrimination power for standard and poor quality materials of the same type are not comparable.
- No single device is capable to detect all the designed differences between the standard and poor quality materials according to the current test methods. However per asphalt type (PA, BBTM and SMA) specific devices appear capable in detecting the designed differences.
- The test methods have relative large geometric standard deviations (often more than 30%), but the number of slabs tested in this research (4) provides enough potential to discriminate between poor quality and standard materials for a large set of the tested devices.
- No uniform correlation between the devices could be found nor their results could be culled or unified for a particular performance/loading in time that would convert to one common measure.

List of Tables

List of Figures

Table of content

1 Introduction

The trans-national research programme "**Call 2014: Asset Management and Maintenance**" was launched by the Conference of European Directors of Roads (CEDR). CEDR is an organisation which brings together the road directors of 25 European countries. The aim of CEDR is to contribute to the development of road engineering as part of an integrated transport system under the social, economic and environmental aspects of sustainability and to promote co-operation between the National Road Administrations (NRA).

The participating NRAs in this Call are Belgium-Flanders, Finland, Germany, Ireland, Norway, the Netherlands, Sweden, United Kingdom and Austria. As in previous collaborative research programmes, the participating members have established a Programme Executive Board (PEB) made up of experts in the topics to be covered. The research budget is jointly provided by the NRAs who provide participants to the PEB as listed above.

Ravelling (or fretting): the loss of aggregates at the road surface, is a common mode of early failure for many types of asphalt pavement. Recently several simulative laboratory tests have been developed to give an indication of the ravelling potential of an asphalt mixture. These tests use scuffing devices that repeatedly apply a scuffing action to slab samples to replicate in service loading. The test methods for four such scuffing devices were up as a draft technical specification by Comité Européen de Normalisation (CEN) as prCEN/TS 12697-50, Resistance to scuffing. However, these methods needed to be culled or combined so that there is only one (harmonised) test method for this one property before the technical specification can be converted into a test standard.

The CEDR-DRaT project looked at the methods of test and the results produced for the four scuffing devices in order to identify:

- The extent to which sample preparation needed to be standardised, such as compaction level, evenness, storage conditions and age when tested.
- The most effective method of measurement in terms of extent of differentiation, validity as a measure of ravelling and practicality.
- Whether the results from one or more scuffing devices could be validated from experience on site.
- Whether the results from different scuffing devices could be converted to a common measure.
- Estimates of the precision of the results with each scuffing device or, if the results could be converted to a common measure, of the common measure.
- Whether the results from either pair of similar devices were comparable and their results were reproducible.
- A procedure to identify if other scuffing devices could be used for the standard test.

These findings might be the same for all asphalt mixture types or different for different types.

The evaluations were based on three variants of three asphalt mixture designs that were tested using six scuffing devices (two of two of the devices and one each for the other two devices) with four replicates of each combination of mixture, variant and test device. All testing was undertaken on laboratory prepared samples. The validation of the test methods was sought by identifying how mixtures with each tested mix design performed on site or in trials.

The overall objective was to provide advice on how to refine prCEN/TS 12697-50 to become an acceptable standard with a draft incorporating that advice.

The CEDR-DRaT project was organized in 5 Work Packages (WPs; see fig. 1). The test slabs were prepared by one single laboratory in WP2 'Sample preparation' and were sent to the different labs for testing in WP3 'Test program'. The output of WP3 was the collection of test results from the individual labs, which was used as input to WP4 'Analysis'.

This report describes the execution and outcome of WP4 'Analysis'.

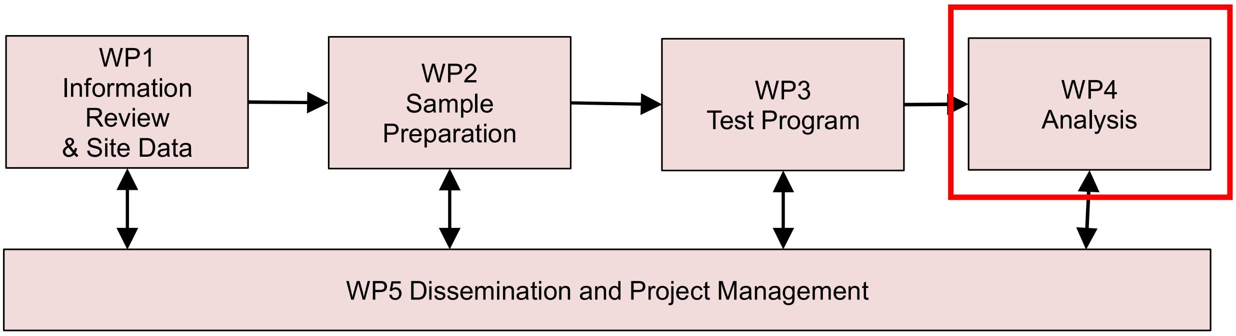


Figure 1: Project organization

2 Description of WP4

2.1 Objectives

The objectives of WP4 ‘Analysis’ were:

Performing the analyses of the measurement data obtained in order to determine:

- the precision of the results with each scuffing device and the similarity (reproducibility) of the results from either pair of similar scuffing devices
- the potential to correlate or unify the results of the different scuffing devices to a common measure
- the potential of the different scuffing devices to discriminate between asphalt qualities

2.2 Approach

2.2.1 Precision (repeatability and reproducibility)

The collected individual results from each device were used to determine the variation of similar slabs under near-homogeneous conditions (repeatability) of the test for the different asphalt types. Statistical techniques were employed to find potential outliers. The outcome of the statistical analysis permits the calculation of the effect size producing a significant result as a function of the number of replicate samples.

The results from the two ARTe and the two DSD devices were used to check reproducibility of these devices, which is the variation of similar slabs under heterogeneous conditions.

2.2.2 Correlation between devices

The gathered results of the six different devices were compared over the total time of testing (loading), with measurement data over 5 time intervals. In order to make this comparison, the individual reported measurements at the different intervals were studied. The measured degree of ravelling was compared by the number of cycles/time to reach a similar ravelling condition. The whole shape (or rate) of ravelling development in time/as function of loading was evaluated likewise.

Based on the gathered data, it was investigated whether a correlation of the devices could be specified in general or for each investigated asphalt type. This showed if any correlation between the different devices existed, whether it was asphalt mixture dependent.

2.2.3 Discrimination

Using the data collected for the different asphalt materials, with variation in standard mixtures and low-temperature-compaction/low-bitumen-designed mixtures, the power to detect significant differences between the test results of the scuffing devices was established.

Finally, it was checked whether each device’s ability to discriminate between good and bad mixtures depends on the asphalt mixture.

3 Background and scope of the Analysis

The scuffing devices under investigation have been developed in order to give an indication of the theoretical potential of an asphalt mixture to ravel. These devices use methods that repeatedly apply a scuffing action to slab samples so as to simulate in service loading. The test methods account for four of such scuffing devices:

- the Aachener Ravelling Tester (ARTe),
- the Darmstadt Scuffing Device (DSD),
- the Rotating Surface Abrasion Test device (RSAT) and,
- the Triboroute device.

These 4 devices have been recorded in the draft technical specification by the Comité Européen de Normalisation (CEN). However, these methods were to be evaluated in a round robin test programme to make out whether they can be used interchangeably and, if not, which of them gives the most informative results.

Six partners of the DRaT consortium are regular users/owners of one of the scuffing devices that are to be compared. In particular:

- BAM and ISAC each own an ARTe,
- BRRC and TUD each own a DSD,
- Heijmans owns an RSAT and
- IFSTTAR owns a Triboroute device.

The evaluation of the scuffing devices is based on the testing of three asphalt types: i.e. PA, BBTM and SMA. For each of the types three variants (qualities) are included. The round robin test further involved the scuffing devices of the six consortium partners owning such a device. All testing is conducted on laboratory prepared samples.

TRL nor TNO owns a scuffing device so that they provide an impartial review of the round robin findings.

The rest of this report accounts for the design and analysis of the round robin test. In Section 4, the experimental design of the round robin test is detailed. Section 5 explains the exact test results that are analysed. In Section 6, the methods used for the statistical analysis of the results are explained. In the Sections 7, 8, and 9 the discriminating power, the precision and the potential to unify the results, are discussed respectively. Finally, Section 10 presents conclusions and recommendations.

4 Experimental design

The round robin test involved 4 types of testing devices and 3 types of asphalt: PA, BBTM and SMA. Each of the asphalt types was addressed in a separate round of slab making, transportation to the respective laboratories, storage, and measuring the proneness of the slabs to ravelling with each of the scuffing devices. So there were in fact three separate round robin tests, one for each type of asphalt.

4.1 Materials

For each of the asphalt types, BAM prepared three sets of 16 slabs each (Jacobs, 2016). One set had a reference mix and normal compaction temperature (reference mix), one set had a composition as the reference mix but a lower compaction temperature (variant 1), and one set had less bitumen than the reference mix, but a normal compaction temperature (variant 2). The qualities of the respective sets were designated 'standard' (s) for the reference mix, 'low temperature' (lt) for variant 1 and 'low bitumen' (lb) for variant 2, respectively. Specification for the mixes and variations are given in Table 1.

The lt is a material compacted at a lower temperature than the standard mix to mimic a poorly compacted material (under lab conditions) and lb is a material containing less binder to mimic a poorly designed mixture (under lab conditions). Although the lt variant was compacted to the same density (as is often the case in lab compaction methods) the lower temperature compaction is known to cause imperfect bonds of the mastic between the aggregates and thereby create flaws/damages, that give the material a higher propensity to ravel. Based on expert judgement and in consensus by the consortium these design parameters for the investigation were chosen as material flaws in order to test the discrimination power of the devices.

Table 1: Mixes and mix variations

Mixture type	Mixture code	Binder	Reference (s)	Variant 1 (lt)	Variant 2 (lb)
PA	PA 16 70/100	70/100	Compaction at 150˚C 5,2% bitumen ±20% air voids	Compaction at 105˚C	4,2% bitumen
BBTM	BBTM 6 50/70	50/70	Compaction at 160˚C 5,6% bitumen ±12-19% air voids	Compaction at 110˚C	4,6% bitumen
SMA	SMA 11 PMB	PMB 25/55-55 with 3% SBS polymer	Compaction at 155˚C 6,8% bitumen ±3% air voids	Compaction at 105˚C	5,5% bitumen

Further details about the mix-designs and preparation of the slabs can be found in Deliverable 6 (D6) of the project (Jacobs, 2016).

4.2 Testing devices

Table 2 shows the testing devices owned by each of the six laboratories. The testing procedures with each scuffing device were performed as described in the annexes of pre-norm **prCEN/TS 12697-50**. The specific annexes are indicated in the table.

Table 2: Testing devices of the six laboratories and pre-norm annexes specifying the testing procedure

Laboratory	Device	Annex in prCEN/TS 12697-50
ISAC – TU Aachen	ARTe	Annex A
BAM	ARTe	Annex A
TU Darmstadt	DSD	Annex B
BRRC	DSD	Annex B
Heijmans	RSAT	Annex C
IFSTTAR	TRD	Annex D

The testing details, as agreed on within the DRaT project, are described in Deliverable 7 (D7) of the project (De Visscher, 2017).

For the purpose of the analysis of the round robin test as performed in the DRaT project (with the three subsets of 3 materials), the scuffing devices included in the test can be divided into two categories:

- Devices that work with quarter slabs; 'Q-devices'
- Devices that work with whole slabs: 'W-devices'

There are three laboratories with Q-devices (TUD, BRRC, IFSTTAR) and three laboratories with W-devices (BAM, Heijmans, ISAC). Each set of 16 slabs of the same quality and asphalt type was randomly divided into four subsets of four slabs. Random distribution was decided in order to minimise the effect of any possible between-slab-variation as an effect of the production sequence. The laboratories with W-devices hence each received one randomised subset. The slabs in the remaining subset of four slabs were cut into quadrants. Each laboratory with a Q-device received a randomised quadrant (in order to minimise the effect of any possible within-slab-variation) from each of the slabs in that subset. Four quadrants were not used for the scuffing tests.

5 Primary and secondary test results

Each of the participating laboratories provided data on a total of 36 asphalt slabs. Upon arrival, the slabs or quarter-slabs were stored for some time. They were subsequently tested in the laboratory's scuffing device. All six laboratories report:

1. The weight of the slab or quarter-slab prior to scuffing.
2. The weight of the slab or quarter-slab at the end of the scuffing test.

The weight loss of the slabs is calculated by subtracting the weight after scuffing from the weight prior to scuffing. This represents the total loss of mass of the test piece. This includes aggregate loss but also the loss of asphalt mortar present at the sample surface.

These weight losses are referred to as the **PRIMARY** test results.

Note: IFSTTAR reported weight losses for PA based on 10 kilocycles, for BBTM based on 6-10 kilocycles and for SMA based on 12-14 kilocycles. The primary outcome for this laboratory was taken to be the weight loss after 10, 6 and 12 kilocycles for PA, BBTM and SMA, respectively.

As the current pre-norm does not fully define how the mass-loss should be determined, also a secondary outcome variable is considered:

- For TUD, BRRC and Heijmans, the weight loss measured by a direct weighing of the material removed by the scuffing is considered. The material removed is either all material that comes from the surface (aggregate and mortar – TUD and BRRC) or could be a selection of the material removed (aggregates > 2 mm - Heijmans)
- For IFSTTAR, the weight loss of PA and SMA slabs after 6,000 cycles only is considered.
- For BAM, additional weight losses that were based on weights including a wooden frame used for the testing are considered.
- For ISAC, there was no obvious secondary outcome variable.

The alternative weight loss measurements are referred to as the **SECONDARY** test results.

The data file used for all statistical analyses is reproduced in Appendix 1.

6 Statistical analysis methods

6.1 t-Tests

Separate statistical analyses are conducted for each type of asphalt and each laboratory. The purpose of each analysis was to establish whether the weight loss in the low temperature (lt) and low bitumen (lb) sets of slabs is higher than in the standard (s) set. (The devices for which this applies are therefore able to find low quality mixes). The simplest way to reach this purpose is to conduct one-sided t tests on the differences between the low temperature and low bitumen sets with the standard set. Based on the variation of the weight losses within each of the three sets, such a t test results in a probability that the difference between a test set of slabs (low temperature or low bitumen) and the standard set can be larger than the observed value merely by chance. This probability is called a P value. If the P value is small, say, $P < 0.05$, the difference is unlikely to occur merely by chance, and it is said that a statistically significant difference has been established.

The t test results in correct P values if the following assumptions are met:

1. Individual weight losses in each set of slabs are normally distributed.
2. The standard deviation of the data is the same for each set of slabs.

It is assumed that the data for each laboratory and asphalt type had a constant relative standard deviation rather than a constant absolute standard deviation, because data near zero are generally less variable than those further away from zero. It can be shown that data with a constant relative standard deviation on the original scale can be turned into data with a constant absolute standard deviation by a log transformation. For this reason, the actual statistical analysis was carried out on log-transformed data.

The reports of the six laboratories showed deviations from the measurement in three cases. The results of these measurements were disregarded in the data analysis. An inspection of TNO of the datasheets sent in by these laboratories identified three further possible outlying weight losses in the data. It is statistically tested whether these weight losses could indeed be considered as outliers by fitting a statistical regression model that includes indicator variables for the three asphalt qualities and a further indicator variable for the outlier. The regression coefficient for the outlier indicator variable may or may not differ from zero in statistically significant way. If it does, the corresponding weight loss is removed from the dataset.

6.2 Data analysis flow

The data analysis flow is summarized for each of the laboratories and each of the asphalt types:

- The results for measurements with deviations from protocol are removed.
- A statistical model is fitted to the weight loss data that include perceived outliers, if any.
- Statistically significant outliers are removed from further consideration.
- t tests are conducted to compare the weight loss in the low temperature and low bitumen groups with the weight loss in the standard group.
- It is checked graphically whether there were any remaining outliers in the data.
- The compatibility is checked graphically with the assumption of a normal distribution and a homogeneous spread of the results.

7 Discrimination; evaluation of quality differences

For each type of asphalt, each of the laboratories reported the results for 36 slabs in total. For the grand total of 216 slab results, laboratory records only suggested irregularities in the measurement procedure for three of the slabs. The corresponding results were removed without further consideration. Inspection of the results suggested three possible outliers. Statistical tests and subsequent inspection of the development of weight loss over time confirmed outlying weight losses in these three cases. The corresponding results were not included in the subsequent analyses. Details on the identification of the slabs whose results were removed are given in Appendix 2.

Summary tables of means and standard deviations after removal of the outliers and irregular measurements are given in Appendix 3. In this section, the graphical summaries of the results are discussed including those of the statistical tests.

7.1 PA results

Figure 2 shows the (primary) weight loss of the PA slabs calculated by subtracting the slab weight after scuffing from the weight prior to scuffing.

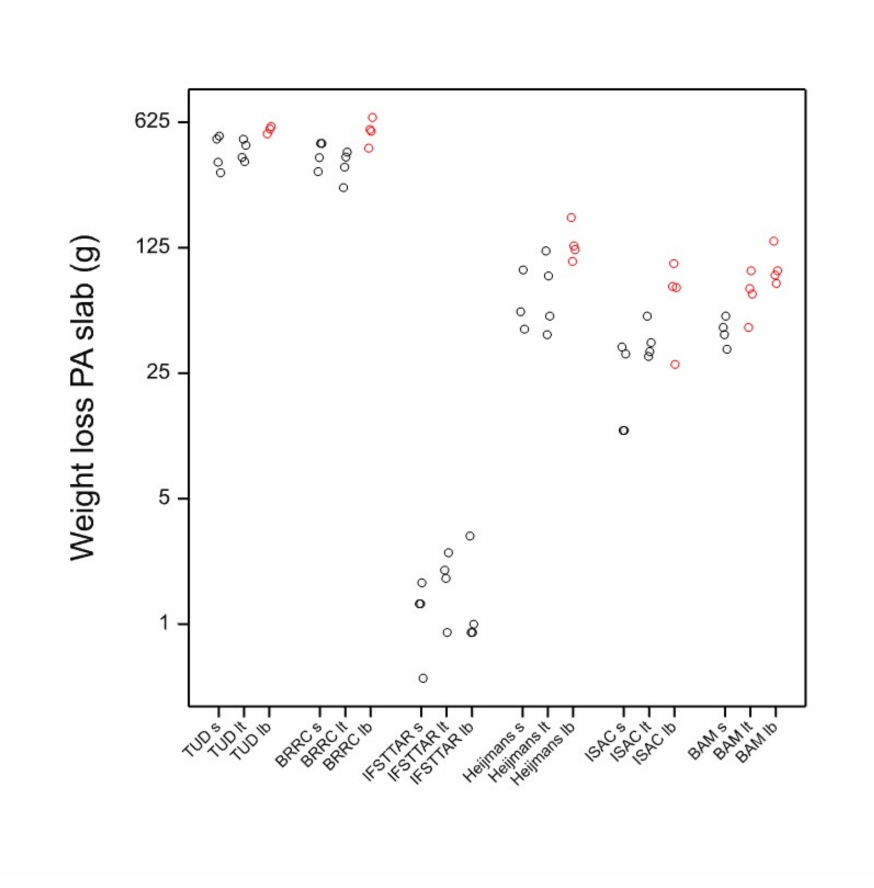


Figure 2: Primary weight loss results for PA

The vertical axis expresses the weight loss in gram. The horizontal axis shows 18 groups, defined by the laboratory and the quality of the asphalt (s: standard mix, lt: mix compacted at a lowered temperature, lb: mix containing less bitumen than standard).

Each circle corresponds to the weight loss of an individual slab. In most cases, the groups include results for four slabs; the groups TUD/lb and Heijmans/s include three slabs; the omitted results have been removed in view of deviations of the measurement protocol.
For each laboratory, a separate statistical analysis was conducted. Whenever a low temperature group or a low bitumen group had a significantly higher weight loss than the group of standard slabs, the weight losses are given in red.

A striking feature of the data displayed in Figure 2 is the big difference in weight losses between the different scuffing devices or laboratories. The weight losses for IFSTTAR are all less than 5 gram, while those from TUD and BRRC are around 500 gram. Weight losses for the remaining laboratories are somewhere between these extremes.

A second notable feature of the figure is that all scuffing devices excepting IFSTTAR's Triboroute can detect differences between the low-bitumen group and the standard.

Third, in the primary weight loss results, only BAM's ARTe revealed an increased weight loss for the low temperature slabs.

Figure 3 shows the alternative (secondary) weight loss results for TUD, BRRC, IFSTTAR, Heijmans and BAM; there were no obvious alternative weight loss measures for ISAC.

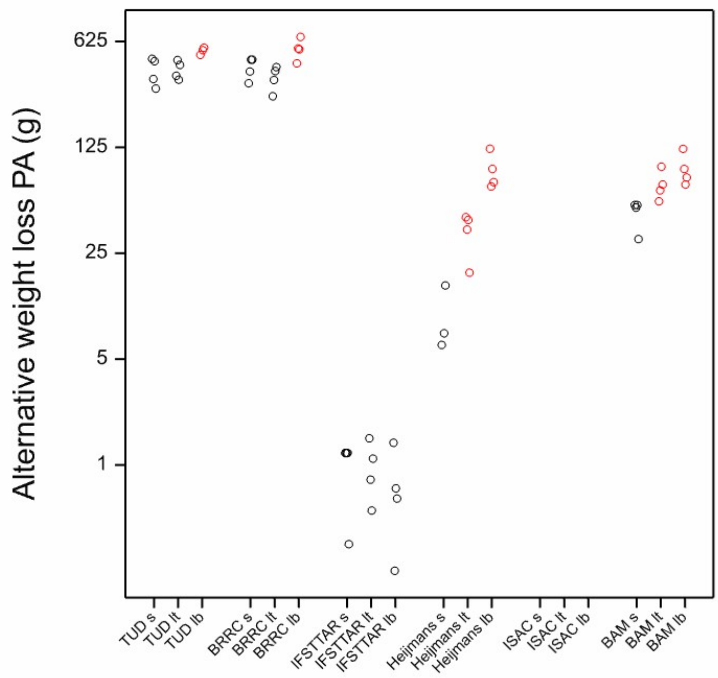


Figure 3: Alternative (secondary) weight loss results for PA

Interestingly, the secondary weight loss results reveal an increased weight loss in low temperature slabs for Heijmans, where the primary results did not reveal such an increase. This is not due to a decrease in variability among the slabs of a group, but rather to the alternative weight loss being somehow more informative.

Figure 2 and Figure 3 further show that the DSD devices of TUD and BRRC give very similar results. The ARTe results of ISAC and BAM show more difference, with BAM reporting higher weight losses on average.

7.2 BBTM results

The results for BBTM are shown in Figure 4. The upper panel of the figure shows the (primary) weight loss of the BBTM slabs calculated by subtracting the slab weight after scuffing from the weight prior to scuffing. The lower panel shows the alternative (secondary) weight loss results for TUD, BRRC, Heijmans and BAM (no obvious alternative weight loss measures for ISAC for any type of asphalt; IFSTTAR has no secondary outcome for BBTM).

Most of the 18 groups defined by the laboratory and the quality of the asphalt include results of four slabs. The group ISAC/lt includes three slabs; the omitted results have been removed in view of deviations of the measurement protocol. The groups IFSTTAR/lt, IFSTTAR/lb and ISAC/lb include three slabs as well; the omitted results have been identified as outliers.

It is observed from the test results that the overall weight loss compared to the results from the PA slabs is much lower. The weight losses for almost all laboratories are close to a factor of 5 lower under the same testing conditions.

The statistical analysis procedure for the BBTM results was the same as the procedure for the PA results. The Triboroute of IFSTTAR and the ARTe of BAM revealed a significantly increased weight loss of the lb slabs when compared with the standard slabs. The weight loss increase for the lt groups was not greater than for the standard groups for any of the laboratories. However, for IFSTTAR, there was a statistically significant decrease in weight loss for the lt slabs ($P = 0.002$). This observation is the only established contradiction to the premise of the designed quality difference of the lt and lb slabs. This is not explicitly marked in the figure, since the primary interest is in detecting increases in weight loss.

The results for BBTM, like the PA results, showed substantial differences among the various scuffing devices. The DSD devices of TUD and BRRC give similar results. The ARTe results of ISAC and BAM are different, with BAM again reporting higher weight losses on average. The discrimination power of the devices and test methods apparently differs depending on the material type.

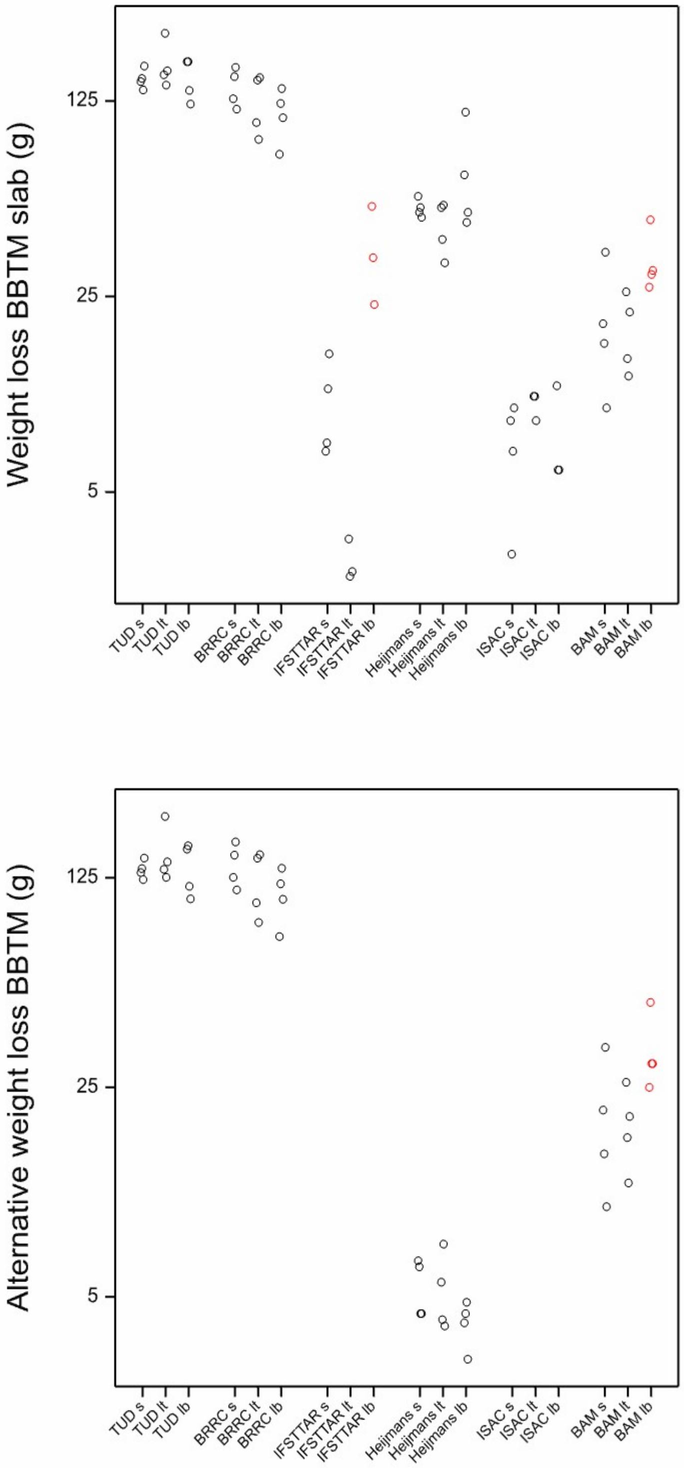


Figure 4: Primary weight loss for BBTM (upper panel) and alternative (secondary) weight loss for BBTM (lower panel)

7.3 SMA results

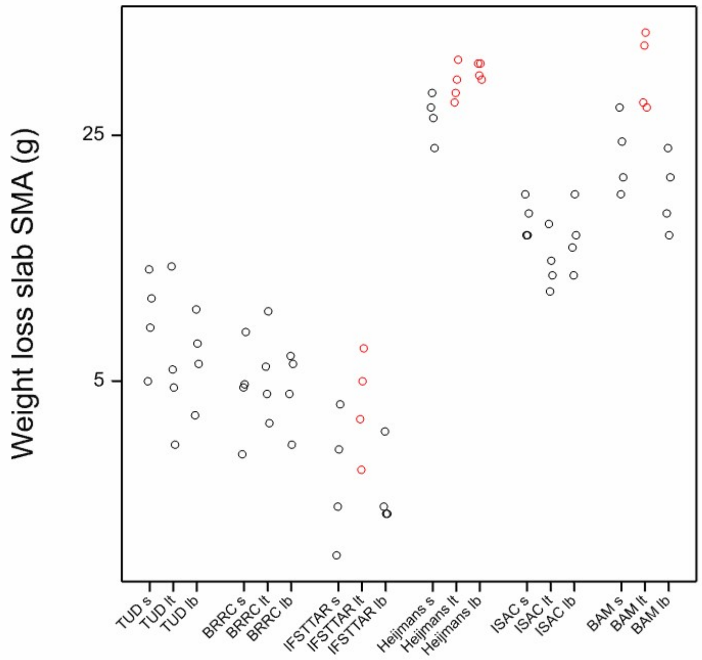
The results for SMA are shown in Figure 5. The upper panel of the figure shows the primary weight loss of the SMA slabs calculated by subtracting the slab weight after scuffing from the weight prior to scuffing. The lower panel shows the alternative (secondary) weight loss results for TUD, BRRC, IFSTTAR, Heijmans and BAM (no obvious alternative weight loss measures for ISAC. All of the 18 groups defined by the laboratory and the quality of the asphalt include results of four slabs.

It is observed from the test results that the overall weight loss compared to the results from the PA and BBTM slabs is very low. The weight losses for almost all laboratories do not exceed the 25 g limit after the full testing procedure. Hence, perhaps it is difficult to speak of ravelling for most of the devices. This indicates that either the current test procedures for the larger part of the devices are too limited to cause ravelling to the material or/and the designed SMA material is unsusceptible to ravelling damage in general.

The statistical analysis procedure for the SMA results was the same as the procedure for the PA and BBTM results.

- The results of TUD, BRRC and ISAC did not reveal differences among the three SMA qualities.
- The primary weight loss results for IFSTTAR were based on 12,000 cycles. These results showed a statistically significant higher weight loss for the lt group when compared with the standard group. The secondary results for IFSTTAR were based on 6,000 cycles. Interestingly, these results showed a statistically significant higher weight loss for both the lt group and the lb group when compared with the standard group.
- For Heijmans, the primary weight loss calculated by subtracting the slab weight after scuffing from the weight prior to scuffing revealed higher results for both the lt group and the lb group when compared with the standard group. The secondary weight loss by direct weighing of the of the material removed by the scuffing revealed higher results for the lb group only.
- For BAM, the primary as well as the secondary weight loss measures showed an increased weight loss for the lt group when compared with the standard group.

The differences among the various scuffing devices for SMA ware much less substantial than for the PA and BBTM asphalt types. The DSD devices of TUD and BRRC give similar results. Unlike the results for PA and BBTM, these results were lower than those for the other scuffing devices. The ARTe results of ISAC and BAM clearly differ, with the ISAC results being on average smaller than those for BAM.



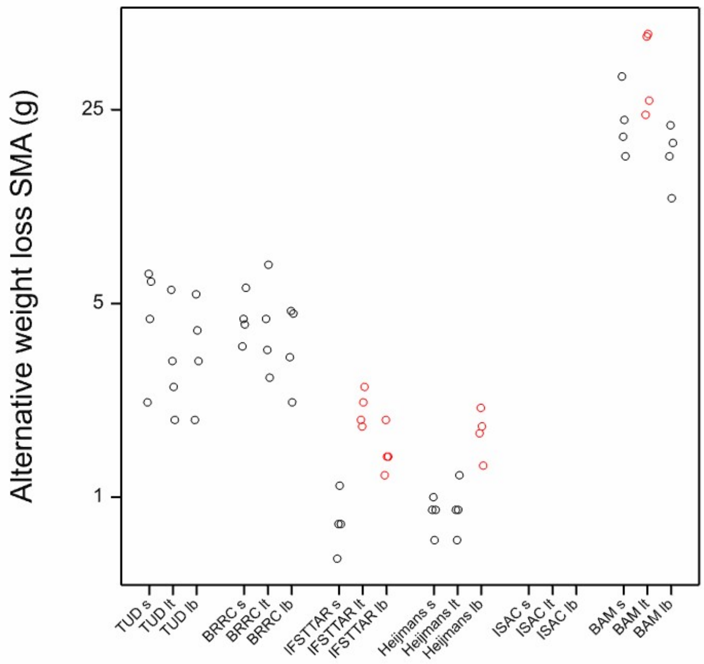


Figure 5: Primary weight loss for SMA (upper panel) and alternative (secondary) weight loss for SMA (lower panel)

7.4 Evaluation per device and laboratory

In this section, the results of the round robin tests are discussed per device and laboratory.

7.4.1 DSD

For both the TUD and BRRC, the results were obtained with a DSD scuffing device. The BRRC results on average were smaller than those for TUD by a factor of 1.2. For both laboratories, the PA results for Ib were higher than the PA results for the standards in a statistically significant way. In fact, this was the only increased weight loss with respect to the standard that could be established for this device within the round robin investigation.

7.4.2 Triboroute

The IFSTTAR results were obtained using a Triboroute device. There was a statistically significant increase in weight loss found with respect to the standard for BBTM / Ib (6,000 cycles), SMA / It (12,000 cycles), SMA / It (6,000 kilocycles) and SMA / Ib (6,000 kilocycles). Apparently, differences between asphalt qualities observed earlier in the test procedure are larger than those later on.

7.4.3 RSAT

The results of Heijmans were obtained with an RSAT device. Using this device, a statistically significant increase in weight loss with respect to the standard was established for PA/It, PA/Ib, SMA/It and SMA/Ib. Curiously, the statistically significant increases in weight loss for PA/It and SMA/It was only observed for one of the two weight loss measures (primary and secondary), and a different one in each of these cases. For SMA/It the result holds in case the weight loss is calculated as the difference in initial and final weight of the tested slabs (primary result). For PA/It, the result holds in case the weight loss is established by direct weighing of the material released by the scuffing (secondary result).

7.4.4 ARTe

The results of both ISAC and BAM were obtained with an ARTe device. However, the results obtained strongly differ among the two laboratories. The ARTe of ISAC only established a statistically significant increase in weight loss for PA/Ib with respect to standard PA. The ARTe of BAM established increased weight losses for PA/It, PA/Ib, BBTM/Ib and SMA/It. The different results for one and the same type of device can be explained only by a different experimental conduct or by (unknown) differences between the individual devices.

8 Precision; variability among slabs from the same lot

8.1 Standard deviation among slabs and statistical power

The comparisons of the three qualities of each type of asphalt were conducted using a statistical test procedure. One of the elements in the procedure is the calculation of the standard deviation based on the sets of four or three slabs of one and the same lot. The smaller the value of the standard deviation, the easier it is to detect small increases in weight loss. It can be shown that the standard deviation of log-transformed measurements approximates the coefficient of variation of the untransformed measurements. Approximate coefficients of variation, expressed as a percentage of the mean, are shown in Figure 6. The approximations were calculated by taking the variance of the log-transformed primary weight losses for the three qualities of a given mix, averaging the three variances, taking the square root of the result, and multiplying by 100 to obtain a percentage (for logarithms to base 10, the results have to be multiplied by ln 10). Based on the approximate coefficients of variation, the smallest difference in weight loss between test slabs and standard slabs that is detectable with a one-sided t test with a false positive rate of 5% and a probability of detection (power) of 80% was calculated.

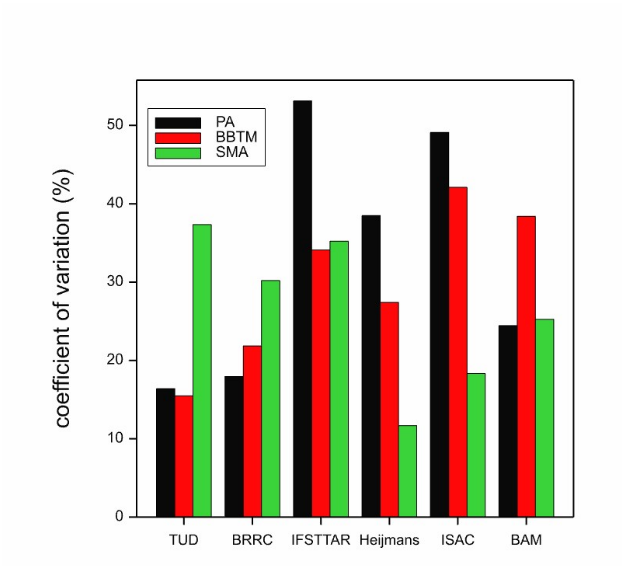


Figure 6: Approximate coefficients of variation in the round robin tests

Results for four, three and two slabs in each of the standard and test groups, expressed as multiplication factors, are shown in Figure 7; see Appendix 4 for the underlying values.

Figure 7 can be interpreted as follows; suppose that TUD is asked to test the weight loss of PA slabs of unknown quality against the weight loss in slabs of an acknowledged good quality using four slabs in each set. A difference in weight loss will be picked up if the true value of the weight loss in the test set is at least 1.4 times higher than the true value in the good-quality set. This is indicated by the height of the first black bar in the upper left panel of Figure 7. The figure displays differences in weight loss between test slabs and standard slabs, expressed as multiplication factors, based on the approximate coefficients of variation in Figure 6.

If the standard group and the test group both include 4 slabs, minimal detectable effect sizes are between 1.3 and 2.9 times the value of the standard quality of the respective asphalt types. For 3 slabs in each of the two groups, the minimum detectable effect sizes run up to 3.7. Finally, the case of 2 slabs in each of the groups can lead to minimum detectable effect sizes as large as 8.3.

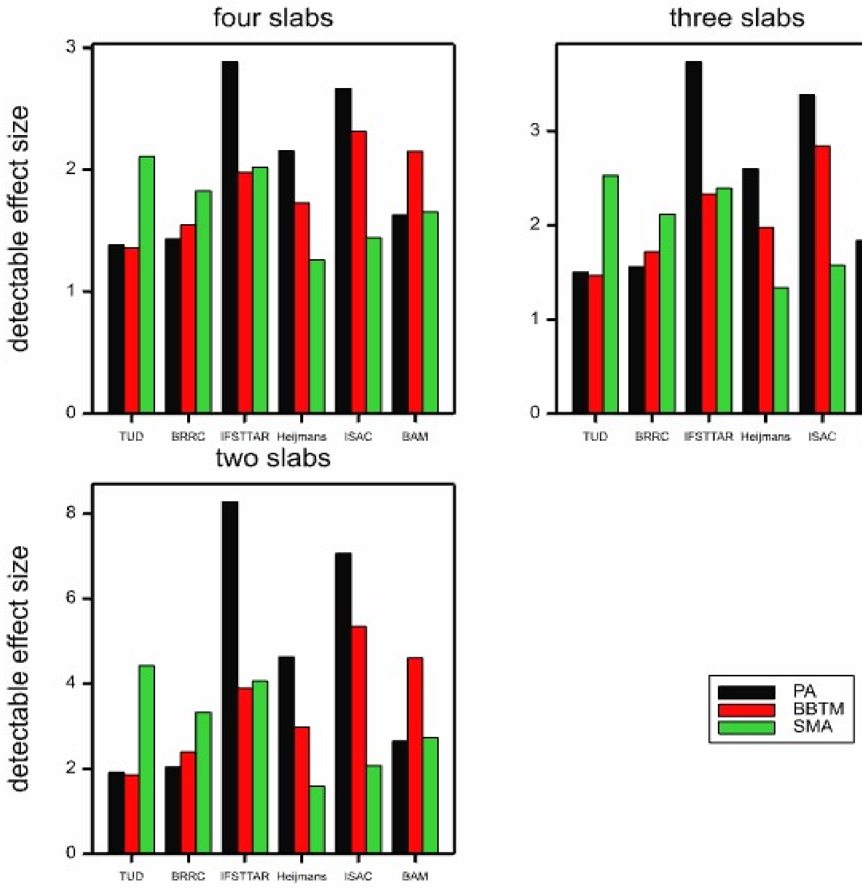


Figure 7: Minimum detectable effect sizes

8.2 Sources of uncontrolled variation

Any test procedure shows variability between tests that are seemingly conducted under identical conditions. For scuffing tests, the following general sources of uncontrolled variation are relevant:

1. The variation among slabs of the same quality.
2. The variation within slabs.
3. The variation introduced be repeatedly applying an experimental procedure.

In the end, regulators may want to set bounds such that no more than a fixed percentage of the slabs or test pieces has a weight loss larger than the bound. Therefore, it is of interest to know each of these sources of variation. However, the variation introduced by repeatedly applying an experimental procedure can be assessed only along with the variation within a slab, because the scuffing is irreversible for the tested surface.

With the test results of TUD and BRRC, it can be checked whether the variation between slabs of the same quality is in line with the variation within the slabs, because the scuffing devices operate on quarter slabs. A statistical mixed model is used to quantify the random variation along with the variation between laboratories and asphalt qualities. For PA and BBTM, there is no indication that different slabs are more variable than different quarters of the same slab. Only for SMA differences could be found and the different slabs account for 37% of the random variation, the remaining 63% is taken up by the variation among quarter slabs, which coincides with repeats of the experimental procedure. These percentages have uncertainties as well. As a result, the stated percentages are at best indicative. However since the weight losses for the SMA in general were only very little and almost no ravelling was induced, it is questionable whether the observed difference between the plates based on the current results is relevant for this comparative study.

8.3 Adjustments for testing conditions

The standard deviations as reported in Section 8.1 per material and in Appendix 3 per mix type quantify the variability among slabs of the same quality tested under conditions that are held constant as far as is feasible. However, in practice these conditions do vary. If the resulting variability can be captured by a measurable property, the results of the slabs can be corrected for the variation in that property so that the remaining slab to slab variation is smaller than before.

8.3.1 Temperature control

Experts on scuffing tests have stressed the impact of the testing temperature on the release of slab material. For this reason it is investigated the effect of the average testing temperature on the weight loss for the three types of asphalt and the six laboratories. The statistical models used for this evaluation included indicator variables for the three asphalt types.

The statistical models checked whether temperature differences *on top of* differences in asphalt quality contributes to the material release. This did not prove to be the case. For BRRC - all three asphalt types, ISAC - BBTM and SMA, Heijmans - BBTM, IFSTTAR - SMA and BAM - SMA, further statistical tests showed that the asphalt qualities were tested at different test temperatures. So, for these cases, if an asphalt quality differs from the standard, this might be due to a difference in test temperature or to the quality itself. However, only a few measurements went outside the set tolerance of ± 2˚C and since the statistical test show that temperature had no significant effect, the tests which stayed within the tolerance should not be doubted.

8.3.2 Testing sequence

Inspection of the test dates of the slabs revealed that, for four of the laboratories, some of the slabs were tested on the same date, while other slabs were tested on other dates. It is thought conceivable that the order of the scuffing test affects the results. To investigate this option, a statistical mixed model is fitted to the data for each asphalt type and laboratory that had more than one slab measured on a day. The model included indicator variables for the asphalt types and a variable that specifies the order of the test within a day.

For the PA results of TUD and BRRC some evidence is found of an effect of the test order. This affects the statistical significance of the comparison between Ib slabs and standard slabs (no significant difference between It slabs and standard).

Table 3 shows the difference between Ib slabs and standard slabs both without and with a correction for measurement order, on a base-10 log scale, while Figure 8 presents the modelled values for standard, It and Ib slabs for first, second and third measurements.

Table 3: Differences between Ib slabs and standard slabs without and with an adjustment for test order.

laboratory	adjustment	Ib minus standard slabs	standard error
TUD	No	0.1264	0.0544
	Yes	0.0784	0.0662
BRRC	No	0.1248	0.0551
	Yes	0.1264	0.0706

For TUD, the difference between Ib slabs and standard slabs decreases upon correction for test order to the extent that the corrected difference is not greater than zero in a statistically significant way. For BRRC, the difference is barely statistically significant after a correction for treatment order.

It is favoured ignoring the possible effect of treatment order in research performed here, because the effect is present in only two out of ten cases where it could appear. Hence, the BRRC results are interpreted as caused by real quality differences rather than as an effect of the treatment order.

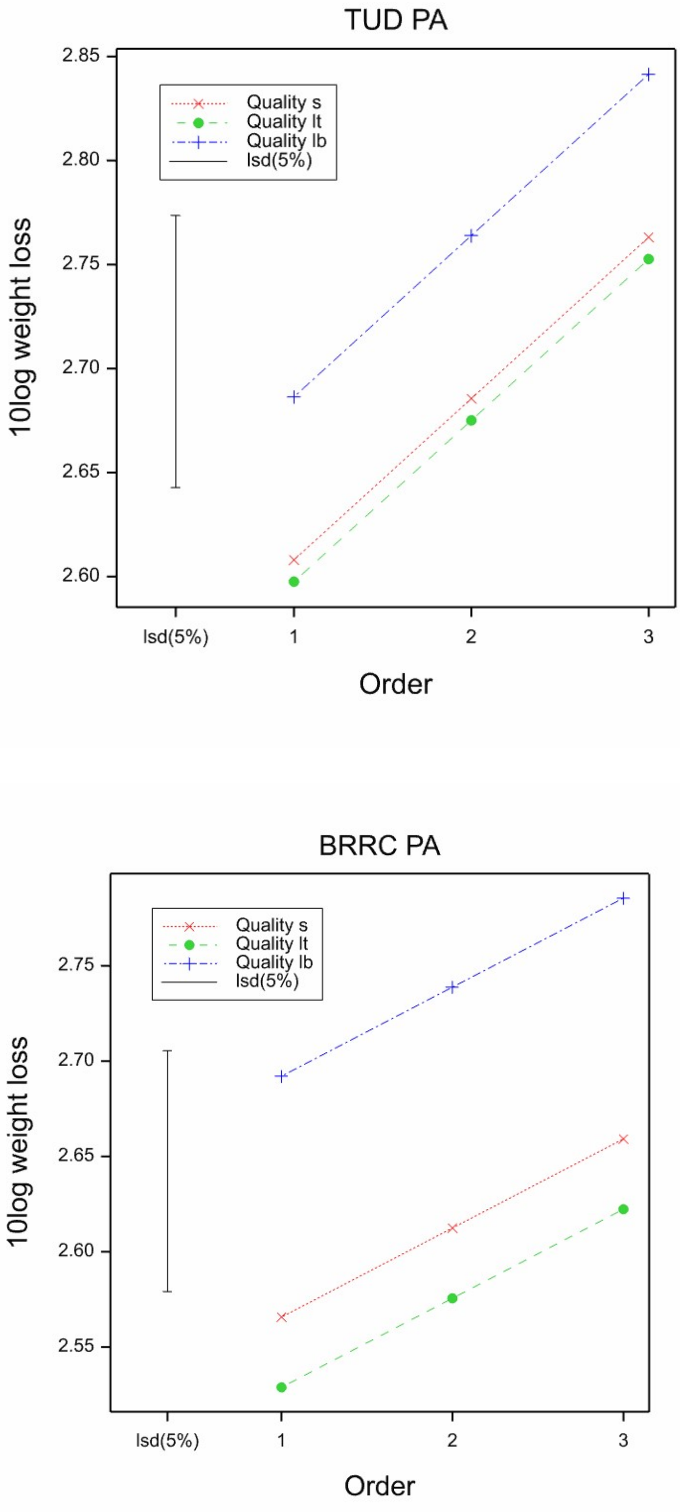


Figure 8: Effect of measurement order for TUD and BRRC, PA slabs.

9 Correlation between devices

Based on the outcome of the DRaT investigation regulators may want to set bounds such that no more than a fixed percentage of slabs or test pieces of a certain material has a weight loss larger than the bound. In order to do so without prescribing a single fixed testing-device it is investigated whether the individual devices can be correlated by finding a scaling factor so that a single bound can be used and whether this factor depends on the asphalt material type.

Comparisons of the final weight losses of the tests were performed in the previous sections. These comparisons may provide evidence on agreement, differences and correlation between the outcomes of the applied test method in combination with the test device. However, they do not necessarily provide evidence on a correlation between the overall behavior of the different devices since only the data points (weight losses) at the start and the end of the procedure were used. Hence, it is interesting to know how the damage develops in the different devices, whether the development differs for the devices or whether the devices show the same pattern of damage development. These issues can be checked, because in the DRaT testing procedure, for each testing of an asphalt slab, intermediate weight loss results were recorded for equally spaced intervals in time (De Visscher, 2017).

In the following analyses, the weight loss development in the various devices were compared by dividing each weight loss at a recorded interval by the final weight loss. So the final result for the primary weight loss of each device is normalised to 1 and earlier weight losses are smaller than 1.

Besides comparing the relative weight losses in time for each device (normalised to the end weight loss value at the end of the test for each device) scaling factors are also calculated to express the weight loss of one device into the weight loss of another device and it is checked whether the scaling factors depend on the asphalt mixture. The scaling factors between two devices were calculated by taking the ratio between two mean weight-loss values at the end of the test for each device. In addition also a scaling factor was calculated compared to an imaginary bound value of 100 g of weight loss.

9.1 PA

Table 4 presents the scaling factors for the PA outcomes.

Table 4: Scaling factors between devices for PA

	TUD	BRRC	IFSTTAR	Heijmans	ISAC	BAM
100 :	0.2	0.2	91.7	1.6	5.0	2.4
TUD :		1.0	388.8	6.9	21.1	10.0
BRRC :			381.4	6.8	20.7	9.8
IFSTTAR :				0.0	0.1	0.0
Heijmans :					3.0	1.4
ISAC :						0.5
BAM :						

The scaling factors are calculated by taking the ratio between the mean values (in grams) of the device belonging to the laboratory in the left column and the (mean)value (in grams) of the device belonging to the laboratory in the top row. For reference the imaginary bound weight loss of 100 g is given in the first row.

The scaling table shows that there is a vast variation between the mutual scaling factors. As was also observed earlier, only the DSD device at TUD and BRRC provide similar scaling factors.

Figure 9 shows the PA weight loss averages for the three material variants (red-s, green-lt, blue-lb) as recorded over time for each of the devices. The data is normalised to the final weight loss of the testing.

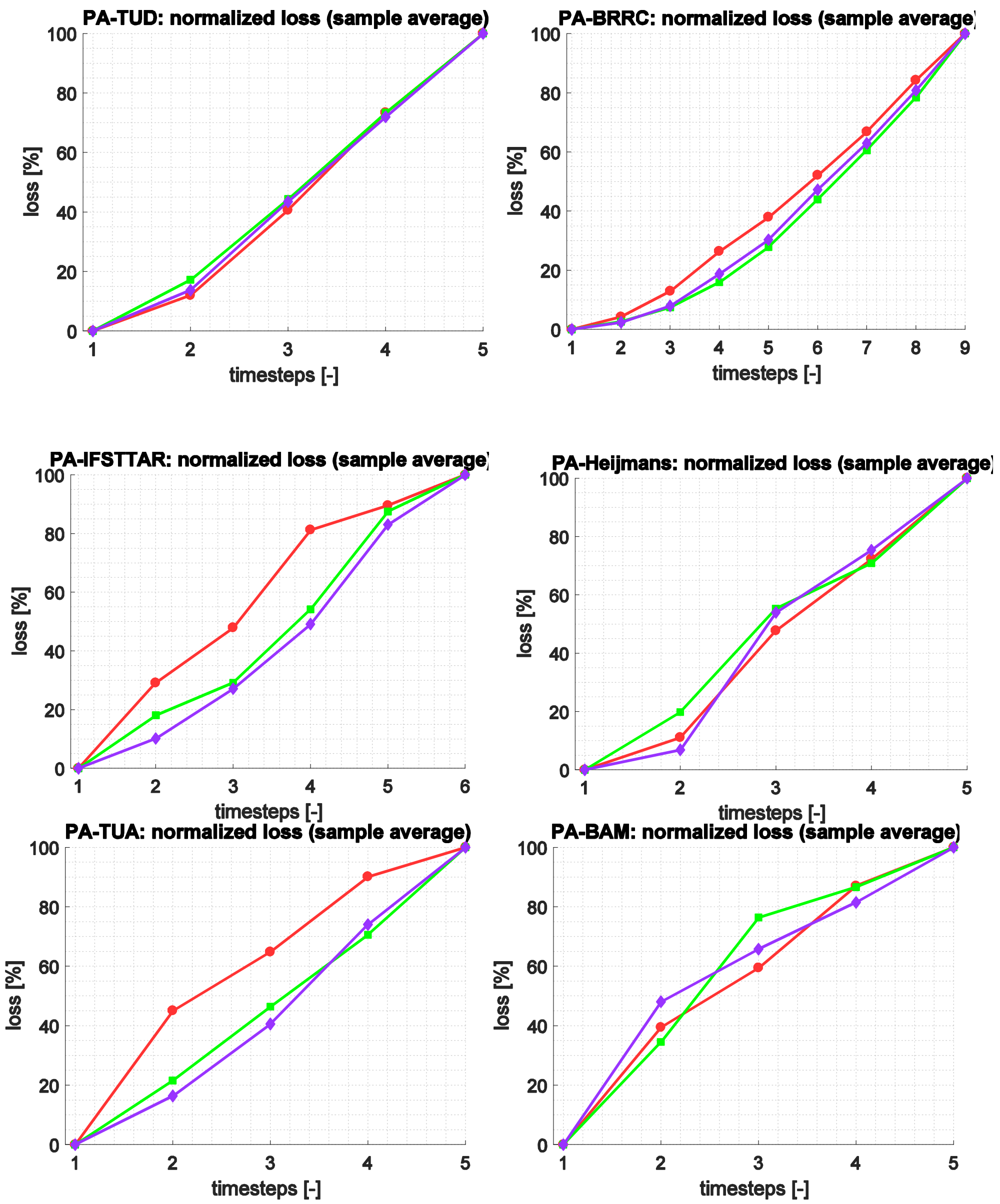


Figure 9: Comparison of damage evolution per device over time for PA

The figure shows for PA that the damage evolution of the DSD device is almost similar for all three material variants and starts off slowly and exponentially increases with the testing time. This behavior is similar for both devices at TUD and BRRC. The Triboroute, RSAT and the ARTe at ISAC all show less consistent behavior for the three materials, but on average demonstrate a more linear trend. The ARTe at BAM shows a more logarithmic development, starting off with a larger damage increment in the early stage of test.

9.2 BBTM

Table 5 presents the scaling factors for the BBTM outcomes.

Table 5: Scaling factors between devices for BBTM

	TUD	BRRC	IFSTTAR	Heijmans	ISAC	BAM
100 :	0.7	0.7	10.1	1.9	15.2	5.3
TUD :		1.1	15.1	2.9	22.7	8.0
BRRC :			14.1	2.7	21.1	7.4
IFSTTAR :				0.2	1.5	0.5
Heijmans :					7.8	2.8
ISAC :						0.4
BAM :						

The table shows that the scaling factors are substantially different from those of PA. Moreover the scaling factors strongly differ mutually, showing a variability for each device in combination with the material type. This renders the use of a single ‘universal’ scaling factor doubtful.

Figure 10 also shows the BBTM weight loss averages for the three material variants (red-s, green-lt, blue-lb) as recorded over time for each of the devices. The data is normalised to the final weight loss of the testing.

The figure shows that the damage evolution for the BBTM materials occurs almost for all devices in a similar fashion in comparison to PA and their mutual differences are still analogously apparent.

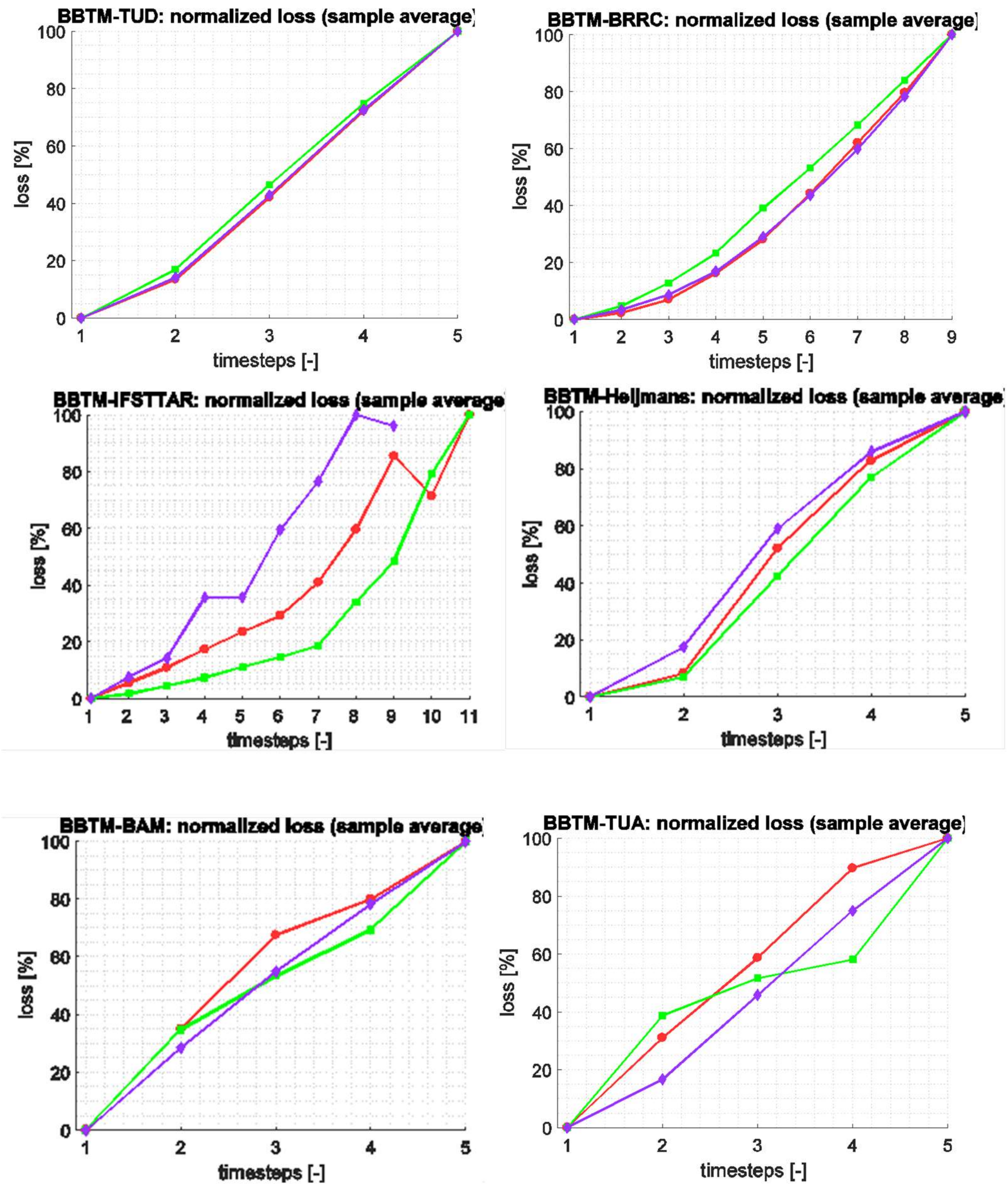


Figure 10: Comparison of damage evolution per device over time for BBTM

9.3 SMA

Table 6 presents the scaling factors for the SMA outcomes.

Table 6: Mutual scaling factors between devices for SMA

	TUD	BRRC	IFSTTAR	Heijmans	ISAC	BAM
100 :	13.3	21.1	37.9	3.5	6.9	4.6
TUD :		1.6	2.8	0.3	0.5	0.3
BRRC :			1.8	0.2	0.3	0.2
IFSTTAR :				0.1	0.2	0.1
Heijmans :					2.0	1.3
ISAC :						0.6
BAM :						

The set of SMA scaling factors confirms the earlier observation (PA and BBTM) that there does not seem to be a constant scaling factor independent of the material type; neither to a fixed reference nor mutually.

Figure 11 shows the SMA weight loss averages for the three material variants (red-s, green-lt, blue-lb) as recorded over time for each of the devices. The data is normalised to the final weight loss of the testing.

In contrast to the earlier figures for PA and BBTM the figure shows that the damage evolution for the SMA materials occurs quite different for the DSD devices and no longer a strong exponential increasing damage trend is observed. For the other devices a similar trend as for PA and BBTM is observed. In specific the characteristic feature of the RSAT and ARTe device remains as was observed for the other materials as well. That is that for the RSAT in the early phase relative low amounts of material are lost, after which the material loss increases strongly in a linear fashion. For the ARTe this behavior is opposite and relative higher amounts are lost in the initial phase after which the material loss still increase but at a lower rate.

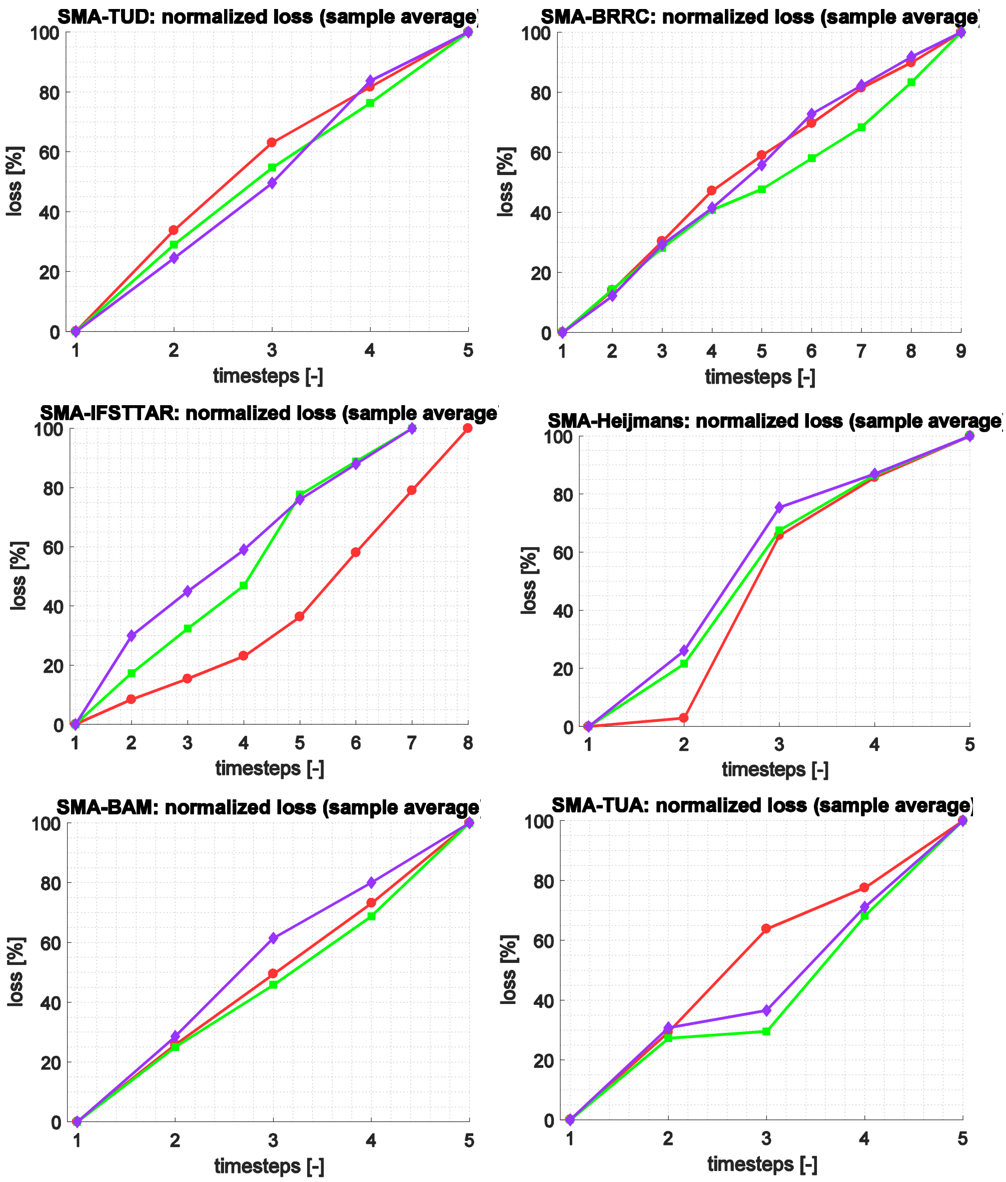


Figure 11: Comparison of damage evolution per device over time for SMA

9.4 Evaluation

The outcome of this comparison between devices and their correlation indicates that there are obvious differences in design and effects of the testing devices. Not only do the devices provide inconstant ratios between the outcomes of their results (scaling factors) for different materials, they also demonstrate a substantial difference in their damage behavior in time. The latter indicates that increasing or shortening of the loading time of the devices could mean stronger differences between the outcomes than yet analyzed here. And such it is considered very difficult to correlate devices uniformly or safely cull and unify their results at a particular performance in time.

10 Conclusions and recommendations

10.1 Discrimination

The purpose of the round robin was to make out whether the scuffing devices can be used interchangeably and, if not, which of them gives the most informative results. To help answering this question, the presence or absence of a statistically significant increase in weight loss between the different qualities of an asphalt type is tabulated in Table 7. Please note that the significance in Table 7 is based on 4 slabs per variant and that the power to detect significant differences is lower in the case of fewer slabs.

Table 7: Statistically significant differences in the round robin tests

Laboratory/ Device	Asphalt type					
	PA		BBTM		SMA	
	(lt)	(lb)	(lt)	(lb)	(lt)	(lb)
TUD / DSD	0	1	0	0	0	0
BRRC / DSD	0	1	0	0	0	0
IFSTTAR /Triboroute	0	0	0	1*	1	1[2]
Heijmans / RSAT	1[2]	1	0	0	1[1]	1
ISAC / ARTe	0	1	0*	0*	0	0
BAM / ARTe	1	1	0	1	1	0

[1]Only for primary weight loss

[2]Only for secondary weight loss

* These results contain the removal of statistical outliers. For the Triboroute BBTM lb variant significance has only been established after removal of an identified statistical outlier.

When comparing the devices per asphalt type:

- For PA, all devices except the Triboroute show an increased weight loss for lb slabs. However, only two of the devices show an increased weight loss for the lt slabs. This suggests that there was indeed a quality difference, which, unfortunately, was not picked up by four of the devices.
- For BBTM no increased weight loss for the lt slabs was established. This could be caused simply by an unintended good quality of these slabs. Two devices show an increased weight loss for lb slabs. This suggests again that there was indeed a quality differences, which, unfortunately, was not picked up by four of the devices.
- For SMA, there were quality differences with respect to the standard in both lt and lb slabs. These were not picked up by all the devices. Based on the overall results it should be commented that current test procedures for the devices are limited to cause raveling to the material or/and the designed SMA material is quite unsusceptible to raveling damage in general.

When looking at the current performances of the devices:

- Roughly, the DSD does not appear sensitive for detecting differences between the different qualities of BBTM and SMA. This device was further not sensitive enough to pick up increased weight loss for PA/lt.
- The Triboroute seems sensitive for BBTM and SMA to detect designed quality differences, but the detection is not consistent for lb and lt differences. It is recommended to standardize the number of cycles for the test, because for different materials different loading times were performed.
- The RSAT does not appear sensitive for detecting differences between the different qualities of BBTM, but it does a good job for PA and SMA, although there are inconsistencies in the primary and secondary weight loss results.
- There are substantial differences between the two ARTe devices. The BAM device seems particularly capable of discriminating between the PA qualities whereas for ISAC, the set of PA/lb slabs was the only set showing statistically significant differences with the standard. Since the different results for one and the same type of device can be explained only by a different experimental conduct or by (unknown) differences between the individual devices (that are currently not fixed in the pre-norm), it is recommended to come to a joint experimental procedure or potentially adapt differences in the mutual design of the devices.

From experience in practice it was anticipated that the tested mixtures would demonstrate an order in raveling propensity following PA > BBTM > SMA, under the same loading conditions. This anticipation is in agreement with the obtained results of all devices, with IFFSTAR's Triboroute being the only exception, as it ranks the PA material last. Although this shows that most devices discriminate between the different asphalt types, since the mutual differences in properties between the materials (PA, BBTM and SMA) are very large, these differences could hardly be marked as the desired level of discrimination power for the devices.

Furthermore, the test results demonstrate that the overall weight loss for the SMA slabs is very low and hardly any physical raveling was actually observed after completion of the tests. So it is questionable whether these results are relevant for qualification of the devices. Perhaps the current test procedures for the largest part of the devices are too limited to cause raveling to the material or/and the designed SMA material is quite unsusceptible to raveling damage in general. Hence, it is recommended to consider adaptations to standardize the test procedures as function of asphalt type (e.g. duration, load). Another recommendation is to supplement specifications for the method on how to establish the raveling, i.e. how to determine the loss-of-weight (e.g. primary vs. secondary results) and how to determine that the weight-loss is an actual raveling effect.

It is concluded that the scuffing devices cannot be used interchangeably because the devices' discrimination potential for standard and poor quality materials of the same type are <u>not</u> comparable.

In addition, it is concluded that <u>no</u> single device is capable to detect all the designed differences between the standard and poor quality materials according to the current test methods. However per asphalt type (PA, BBTM and SMA) specific devices appear capable in detecting the designed differences.

10.2 Precision

The devices show fairly large coefficients of variation, with the consequence that the discrimination potential strongly depends on the number of slabs tested. Due to the large variation in the outcomes of the test we strongly recommend that the number of slabs should not be lower than three (3) in order to keep sufficient discrimination power of the devices to detect lower quality materials.

It is concluded that the test methods have relatively large coefficients of variation (often more than 30%). The four (4) slabs per asphalt mix tested in this project permitted discrimination between poor quality and standard materials in 14 of the 36 combinations of laboratory and mixture variant.

10.3 Correlation between devices

The comparison between the devices for the three materials demonstrates that there are clear differences in design and effects of the testing devices. It appears impossible to find a uniform scaling factor to turn the results of one device to another since the scaling strongly depends on the material tested. Furthermore it is shown that many of the devices show a different damage development in time. Hence, based on the outcomes of this investigation, the results of the individual devices cannot be converted to a common measure.

It is concluded that <u>no</u> uniform correlation between the devices could be found <u>nor</u> that their results could be culled or unified for a particular performance/loading in time that would convert to one common measure.

11 References

De Visscher, J. (2017). *D7- Factual report on test results.* BRRC.

Jacobs, M. (2016). *D6 - Sample preparation.* BAM Infra.

Appendix 1: Datafile used in statistical analyses

The table below shows the data used in the statistical analyses. There are 12 columns:

1. Type: the type of asphalt (PA, BBTM, SMA).
2. Lab: the laboratory (TUD, BRRC, IFSTTAR, Heijmans, ISAC, BAM).
3. ID: the ID of the slab; Mx-y-z is the zth slab of quality y of asphalt type x (x=1: PA, x=2: BBTM, x=3: SMA, y=1: standard, y=2: lt, y=3:lb).
4. Part: the part of the slab tested in a lab with a Q-device. For laboratories with W-devices, the part is recorded as an ‘a’.
5. Date: the date of the test.
6. Order: the order number of the test in case several tests are conducted on the same day.
7. Temp: the average temperature during the test.
8. Suspect: the outlier status of the record (0: no outlier, 1: removed from analysis in view of irregularity in the test process, 2: removed from analysis because it is a statistical outlier; see Appendix 2).
9. M[1]: the mass of the slab prior to testing.
10. M[2]: the mass of the slab at the end of the test.
11. First: the primary outcome; see Chapter 3.
12. Second: the secondary outcome; see Chapter 3.

Type	Lab	ID	Part	Date	Order	Temp	Suspect	M[1]	M[2]	First	Second
PA	TUD	M1-1-5	d	22-3-2016	0	39.61	0	5830.7	5326.4	504.30	480
PA	TUD	M1-1-7	c	24-3-2016	1	40.96	0	5756.1	5381.8	374.30	352.8
PA	TUD	M1-1-8	b	24-3-2016	2	40.88	0	5876.5	5353.5	523.00	462.9
PA	TUD	M1-1-12	a	29-3-2016	0	40.35	0	5834	5507.1	326.90	305
PA	TUD	M1-2-4	d	31-3-2016	1	39.23	0	5889.2	5491	398.20	370.8
PA	TUD	M1-2-5	c	31-3-2016	2	39.95	0	5834.5	5332.2	502.30	470.8
PA	TUD	M1-2-10	a	7-4-2016	1	39.81	0	5800.9	5423.5	377.40	348.9
PA	TUD	M1-2-13	b	7-4-2016	2	40.15	0	5803	5338.1	464.90	437.7
PA	TUD	M1-3-3	c	11-4-2016	0	39.64	0	5778.7	5240.5	538.20	509.6
PA	TUD	M1-3-11	d	15-4-2016	1	41.29	1	5753.3	5328.4	424.90	327.9
PA	TUD	M1-3-13	a	15-4-2016	2	39.84	0	5754.8	5182.3	572.50	547.5
PA	TUD	M1-3-14	b	15-4-2016	3	39.75	0	5805.8	5214.2	591.60	568.5
PA	BRRC	M1-1-5	b	24-3-2016	0	41.55	0	5851.5	5519.6	331.90	**330.75**
PA	BRRC	M1-1-7	a	25-3-2016	1	41.42	0	5806	5409.1	396.90	**395.75**
PA	BRRC	M1-1-8	d	25-3-2016	2	41.36	0	5877.8	5401.4	476.40	473.6
PA	BRRC	M1-1-12	c	25-3-2016	3	41.08	0	5812.4	5336.5	475.90	474
PA	BRRC	M1-2-4	b	5-4-2016	1	40.31	0	5821.9	5551.6	270.30	271.7
PA	BRRC	M1-2-5	a	5-4-2016	2	40.12	0	5903.9	5552.3	351.60	347.6
PA	BRRC	M1-2-10	c	7-4-2016	1	39.81	0	5857.9	5458.5	399.40	398.4
PA	BRRC	M1-2-13	d	7-4-2016	2	40.14	0	5913.5	5487.4	426.10	423.3
PA	BRRC	M1-3-3	a	18-4-2016	0	40.41	0	5853.9	5405.9	448.00	447.8

PA	BRRC	M1-3-11	b	19-4-2016	1	40.54	0	5866.1	5297	569.10	561.4
PA	BRRC	M1-3-13	c	19-4-2016	2	40.59	0	5852.3	5295.7	556.60	553.7
PA	BRRC	M1-3-14	d	19-4-2016	3	40.31	0	5788.3	5123.9	664.40	669.7
PA	IFSTTAR	M1-1-5	a	24-3-2016	0	21.00	0	3745.3	3744	1.30	1.2
PA	IFSTTAR	M1-1-7	d	21-4-2016	0	21.50	0	3568.8	3567.5	1.30	1.2
PA	IFSTTAR	M1-1-8	c	29-3-2016	0	19.50	0	3643.1	3641.4	1.70	1.2
PA	IFSTTAR	M1-1-12	b	31-3-2016	0	21.20	0	3594.8	3594.3	0.50	0.3
PA	IFSTTAR	M1-2-4	a	4-4-2016	0	20.50	0	3599	3597	2.00	1.5
PA	IFSTTAR	M1-2-5	d	5-4-2016	0	19.70	0	3716	3714.2	1.80	0.8
PA	IFSTTAR	M1-2-10	b	8-4-2016	0	20.00	0	3645.8	3644.9	0.90	0.5
PA	IFSTTAR	M1-2-13	c	12-4-2016	0	20.50	0	3597.2	3594.7	2.50	1.1
PA	IFSTTAR	M1-3-3	d	14-4-2016	0	19.00	0	3548.8	3545.7	3.10	1.4
PA	IFSTTAR	M1-3-11	a	19-4-2016	1	20.80	0	3609.1	3608.2	0.90	0.2
PA	IFSTTAR	M1-3-13	b	19-4-2016	2	20.50	0	3649.1	3648.2	0.90	0.7
PA	IFSTTAR	M1-3-14	c	22-4-2016	0	20.00	0	3565.2	3564.2	1.00	0.6
PA	Heijmans	M1-1-6	a	12-4-2016	0	20.50	0	16730	16675	55.00	6.2
PA	Heijmans	M1-1-9	a	13-4-2016	0	19.75	1	16556	16405	151.00	80.3
PA	Heijmans	M1-1-11	a	26-4-2016	0	20.38	0	16746	16652	94.00	7.4
PA	Heijmans	M1-1-18	a	19-5-2016	0	20.00	0	16678	16634	44.00	15.3
PA	Heijmans	M1-2-1	a	29-4-2016	0	19.50	0	16817	16697	120.00	43.2
PA	Heijmans	M1-2-6	a	2-5-2016	0	20.25	0	16866	16825	41.00	35.8
PA	Heijmans	M1-2-9	a	9-5-2016	0	20.00	0	16632	16545	87.00	41.5
PA	Heijmans	M1-2-18	a	23-5-2016	0	20.00	0	16666	16614	52.00	18.6
PA	Heijmans	M1-3-5	a	10-5-2016	0	20.25	0	16445	16261	184.00	122.3
PA	Heijmans	M1-3-6	a	12-5-2016	0	19.75	0	16885	16780	105.00	69
PA	Heijmans	M1-3-7	a	17-5-2016	0	20.25	0	16453	16325	128.00	89.9
PA	Heijmans	M1-3-15	a	18-5-2016	0	20.00	0	16462	16340	122.00	73.6
PA	ISAC	M1-1-4	a	23-3-2016	0	24.00	0	15000	14965	35.00	
PA	ISAC	M1-1-17	a	20-4-2016	0	26.50	0	14683	14671	12.00	
PA	ISAC	M1-1-19	a	19-4-2016	0	25.75	0	14630	14618	12.00	
PA	ISAC	M1-1-20	a	21-4-2016	0	28.00	0	14693	14661	32.00	
PA	ISAC	M1-2-7	a	12-4-2016	0	27.38	0	15259	15207	52.00	
PA	ISAC	M1-2-8	a	11-4-2016	0	26.00	0	14768	14737	31.00	
PA	ISAC	M1-2-16	a	20-4-2016	0	27.75	0	14758	14725	33.00	
PA	ISAC	M1-2-17	a	14-4-2016	0	27.38	0	14892	14855	37.00	
PA	ISAC	M1-3-1	a	12-4-2016	0	27.50	0	14620	14544	76.00	
PA	ISAC	M1-3-4	a	13-4-2016	0	26.50	0	14643	14541	102.00	
PA	ISAC	M1-3-16	a	21-4-2016	0	26.63	0	15041	15013	28.00	
PA	ISAC	M1-3-18	a	22-4-2016	0	29.25	0	14713	14638	75.00	
PA	BAM	M1-1-10	a	24-3-2016	0	21.80	0	21474	21429	45.00	52
PA	BAM	M1-1-13	a	29-3-2016	0	21.13	0	21621	21580	41.00	50
PA	BAM	M1-1-16	a	22-4-2016	0	22.33	0	21544	21492	52.00	52

PA	BAM	M1-1-21	a	25-4-2016	0	21.47	0	21417	21383	34.00	31
PA	BAM	M1-2-2	a	31-3-2016	1	21.20	0	21479	21434	45.00	55
PA	BAM	M1-2-3	a	31-3-2016	2	22.07	0	21605	21531	74.00	65
PA	BAM	M1-2-12	a	7-4-2016	1	20.93	0	21717	21624	93.00	93
PA	BAM	M1-2-14	a	7-4-2016	2	21.93	0	21791	21722	69.00	71
PA	BAM	M1-3-2	a	11-4-2016	0	21.53	0	21414	21278	136.00	122
PA	BAM	M1-3-9	a	13-4-2016	1	21.53	0	21589	21501	88.00	90
PA	BAM	M1-3-10	a	13-4-2016	2	22.13	0	21375	21296	79.00	71
PA	BAM	M1-3-17	a	26-4-2016	0	21.07	0	21437	21344	93.00	79
BBTM	TUD	M2-1-2	d	3-5-2016	1	39.90	0	5707.5	5560.9	146.60	130.1
BBTM	TUD	M2-1-4	c	3-5-2016	2	39.75	0	5780.2	5629.8	150.40	134.2
BBTM	TUD	M2-1-5	b	4-5-2016	1	39.63	0	5761.3	5624.5	136.80	123.3
BBTM	TUD	M2-1-10	a	4-5-2016	2	40.26	0	5839.4	5672.8	166.60	145.5
BBTM	TUD	M2-2-5	d	12-5-2016	1	40.00	0	5801.4	5646.3	155.10	133.4
BBTM	TUD	M2-2-6	c	12-5-2016	2	40.06	0	5857.8	5639.7	218.10	200.2
BBTM	TUD	M2-2-14	c	18-5-2016	1	39.75	0	5779.3	5636.9	142.40	125.4
BBTM	TUD	M2-2-14	d	18-5-2016	2	40.04	0	5778.5	5618.3	160.20	141.3
BBTM	TUD	M2-3-3	c	24-5-2016	0	40.21	0	5747.8	5575.2	172.60	155.7
BBTM	TUD	M2-3-11	d	25-5-2016	1	40.00	0	5753	5580.3	172.70	159.9
BBTM	TUD	M2-3-12	a	25-5-2016	2	40.00	0	5822	5685.7	136.30	117.1
BBTM	TUD	M2-3-13	b	31-5-2016	0	39.93	0	5814	5692.2	121.80	106.3
BBTM	BRRC	M2-1-2	b	3-5-2016	1	40.76	0	5842.9	5715.6	127.30	**125.4**
BBTM	BRRC	M2-1-4	a	3-5-2016	2	40.27	0	5907.1	5754.4	152.70	**148.7**
BBTM	BRRC	M2-1-5	d	9-5-2016	1	41.08	0	5841.1	5676.4	164.70	164.6
BBTM	BRRC	M2-1-10	c	9-5-2016	2	39.20	0	5750.5	5633.6	116.90	113.9
BBTM	BRRC	M2-2-5	b	10-5-2016	1	39.16	0	5921.5	5816.9	104.60	103.2
BBTM	BRRC	M2-2-6	a	10-5-2016	2	39.37	0	5833.6	5685.3	148.30	145.3
BBTM	BRRC	M2-2-11	c	13-5-2016	1	39.46	0	5936.1	5845	91.10	88.7
BBTM	BRRC	M2-2-15	d	13-5-2016	2	39.73	0	5816.5	5665.1	151.40	149.3
BBTM	BRRC	M2-3-3	a	20-5-2016	1	39.39	0	5921	5840.4	80.60	79.6
BBTM	BRRC	M2-3-11	b	20-5-2016	2	39.17	0	5887.4	5765	122.40	119.4
BBTM	BRRC	M2-3-12	c	24-5-2016	1	39.22	0	5797.8	5659.4	138.40	134.5
BBTM	BRRC	M2-3-13	d	24-5-2016	2	39.43	0	5807.4	5698.5	108.90	106
BBTM	IFSTTAR	M2-1-2	a	23-5-2016	0	20.50	0	3652.3	3645.3	7.00	
BBTM	IFSTTAR	M2-1-4	d	26-5-2016	0	20.00	0	3527.2	3519.7	7.50	
BBTM	IFSTTAR	M2-1-5	c	31-5-2016	0	19.50	0	3548	3536.3	11.70	
BBTM	IFSTTAR	M2-1-10	b	2-6-2016	0	20.50	0	3610.6	3595	15.60	
BBTM	IFSTTAR	M2-2-5	a	7-6-2016	0	21.00	0	3581.8	3578.4	3.40	
BBTM	IFSTTAR	M2-2-6	d	9-6-2016	0	21.00	0	3614.2	3611.7	2.50	
BBTM	IFSTTAR	M2-2-11	b	13-6-2016	0	24.00	2	3606	3575.6	30.40	
BBTM	IFSTTAR	M2-2-15	c	17-6-2016	0	22.50	0	3571.7	3569.1	2.60	
BBTM	IFSTTAR	M2-3-3	d	21-6-2016	0	20.50	0	3621.3	3568.8	52.50	

BBTM	IFSTTAR	M2-3-11	a	27-6-2016	0	22.50	0	3584.4	3550	34.40	
BBTM	IFSTTAR	M2-3-12	b	28-6-2016	0	22.00	0	3525.7	3502.3	23.40	
BBTM	IFSTTAR	M2-3-13	c	30-6-2016	0	22.00	2	3492.1	3490.2	1.90	
BBTM	Heijmans	M2-1-3	a	11-7-2016	0	20.63	0	16493	16436	57.00	6.6
BBTM	Heijmans	M2-1-6	a	12-7-2016	0	20.88	0	16403	16353	50.00	6.3
BBTM	Heijmans	M2-1-9	a	13-7-2016	0	21.13	0	16314	16262	52.00	4.4
BBTM	Heijmans	M2-1-15	a	15-7-2016	0	21.13	0	16184	16136	48.00	4.4
BBTM	Heijmans	M2-2-2	a	21-7-2016	0	20.13	0	16719	16667	52.00	5.6
BBTM	Heijmans	M2-2-7	a	25-7-2016	0	20.00	0	16818	16778	40.00	4.2
BBTM	Heijmans	M2-2-10	a	27-7-2016	0	20.13	0	16777	16724	53.00	7.5
BBTM	Heijmans	M2-2-19	a	28-7-2016	0	20.25	0	16965	16932	33.00	4
BBTM	Heijmans	M2-3-5	a	29-7-2016	0	20.38	0	16698	16630	68.00	4.1
BBTM	Heijmans	M2-3-6	a	2-8-2016	0	20.25	0	16582	16468	114.00	4.4
BBTM	Heijmans	M2-3-7	a	3-8-2016	0	20.00	0	16767	16721	46.00	4.8
BBTM	Heijmans	M2-3-14	a	4-8-2016	0	20.38	0	16568	16518	50.00	3.1
BBTM	ISAC	M2-1-1	a	4-5-2016	1	27.63	0	14973	14964	9.00	
BBTM	ISAC	M2-1-14	a	4-5-2016	2	28.25	0	15002	14999	3.00	
BBTM	ISAC	M2-1-16	a	9-5-2016	1	27.00	0	14875	14868	7.00	
BBTM	ISAC	M2-1-17	a	9-5-2016	2	27.38	0	14832	14822	10.00	
BBTM	ISAC	M2-2-8	a	11-5-2016	1	27.25	0	15097	15086	11.00	
BBTM	ISAC	M2-2-9	a	11-5-2016	2	26.75	0	15093	15082	11.00	
BBTM	ISAC	M2-2-17	a	13-5-2016	1	26.25	0	15074	15065	9.00	
BBTM	ISAC	M2-2-18	a	13-5-2016	2	28.63	1	14945	14908	37.00	
BBTM	ISAC	M2-3-1	a	17-5-2016	1	26.88	0	14981	14969	12.00	
BBTM	ISAC	M2-3-4	a	17-5-2016	2	27.75	0	15047	15041	6.00	
BBTM	ISAC	M2-3-15	a	23-5-2016	1	26.75	0	14830	14824	6.00	
BBTM	ISAC	M2-3-17	a	23-5-2016	2	28.88	2	14880	14879	1.00	
BBTM	BAM	M2-1-8	a	9-5-2016	0	22.43	0	24810	24790	20.00	21.0
BBTM	BAM	M2-1-11	a	11-5-2016	0	22.15	0	24809	24792	17.00	15.0
BBTM	BAM	M2-1-13	a	12-5-2016	1	22.28	0	24981	24945	36.00	34.0
BBTM	BAM	M2-1-18	a	12-5-2016	2	22.80	0	25059	25049	10.00	10.0
BBTM	BAM	M2-2-3	a	13-5-2016	0	22.45	0	24953	24927	26.00	26.0
BBTM	BAM	M2-2-4	a	17-5-2016	0	22.88	0	25065	25050	15.00	17.0
BBTM	BAM	M2-2-13	a	19-5-2016	1	22.63	0	25144	25131	13.00	12.0
BBTM	BAM	M2-2-16	a	19-5-2016	2	22.40	0	25182	25160	22.00	20.0
BBTM	BAM	M2-3-2	a	25-5-2016	0	22.23	0	21588	21561	27.00	25.0
BBTM	BAM	M2-3-9	a	26-5-2016	0	22.53	0	24959	24912	47.00	48.0
BBTM	BAM	M2-3-10	a	27-5-2016	0	22.38	0	24870	24840	30.00	30.0
BBTM	BAM	M2-3-16	a	31-5-2016	0	22.15	0	21675	21644	31.00	30.0
SMA	TUD	M3-1-5	d	29-6-2016	0	39.95	0	6800.1	6795.1	5.00	2.2
SMA	TUD	M3-1-7	c	30-6-2016	1	40.58	0	6752.6	6742.2	10.40	6.4
SMA	TUD	M3-1-8	b	30-6-2016	2	40.08	0	6783.4	6776.3	7.10	4.4

SMA	TUD	M3-1-12	a	30-6-2016	3	40.03	0	6675.7	6667.1	8.60	6
SMA	TUD	M3-2-6	d	8-7-2016	0	39.86	0	6872.5	6861.9	10.60	5.6
SMA	TUD	M3-2-7	c	19-7-2016	1	39.69	0	6795.3	6789.9	5.40	3.1
SMA	TUD	M3-2-12	a	19-7-2016	2	40.13	0	6939.6	6934.8	4.80	2.5
SMA	TUD	M3-2-15	b	21-7-2016	0	39.71	0	6811.6	6808.3	3.30	1.9
SMA	TUD	M3-3-3	c	26-7-2016	0	40.19	0	6781.7	6777.7	4.00	1.9
SMA	TUD	M3-3-12	d	28-7-2016	0	40.20	0	6806.3	6798.3	8.00	5.4
SMA	TUD	M3-3-13	a	3-8-2016	1	39.81	0	6808.6	6802.2	6.40	4
SMA	TUD	M3-3-14	b	3-8-2016	2	40.04	0	6801.4	6795.8	5.60	3.1
SMA	BRRC	M3-1-5	b	4-7-2016	0	40.39	0	6868.2	6865.1	3.10	**3.5**
SMA	BRRC	M3-1-7	a	5-7-2016	1	39.88	0	6850.5	6845.7	4.80	**4.4**
SMA	BRRC	M3-1-8	d	5-7-2016	2	40.59	0	6874.5	6869.6	4.90	4.2
SMA	BRRC	M3-1-12	c	25-7-2016	0	40.12	0	6645.5	6638.6	6.90	5.7
SMA	BRRC	M3-2-6	b	25-7-2016	0	40.33	0	6905.1	6899.6	5.50	4.4
SMA	BRRC	M3-2-7	a	26-7-2016	1	40.13	0	6913.4	6908.8	4.60	3.4
SMA	BRRC	M3-2-12	c	26-7-2016	2	40.17	0	6813.6	6805.7	7.90	6.9
SMA	BRRC	M3-2-15	d	26-7-2016	3	40.43	0	6935.1	6931.3	3.80	2.7
SMA	BRRC	M3-3-3	a	28-7-2016	1	39.54	0	6861.3	6856.7	4.60	3.2
SMA	BRRC	M3-3-12	b	28-7-2016	2	39.64	0	6866.4	6860.5	5.90	4.7
SMA	BRRC	M3-3-13	c	29-7-2016	1	38.99	0	6840.7	6837.4	3.30	2.2
SMA	BRRC	M3-3-14	d	29-7-2016	2	39.09	0	6913.5	6907.9	5.60	4.6
SMA	IFSTTAR	M3-1-5	a	1-7-2016	0	22.50	0	4229.2	4227.6	1.60	0.6
SMA	IFSTTAR	M3-1-7	d	4-7-2016	0	20.00	0	4160.1	4157.9	2.20	0.8
SMA	IFSTTAR	M3-1-8	c	4-7-2016	0	20.50	0	4209.6	4206.4	3.20	1.1
SMA	IFSTTAR	M3-1-12	b	5-7-2016	0	22.00	0	4197.2	4192.9	4.30	0.8
SMA	IFSTTAR	M3-2-6	a	18-7-2016	0	24.00	0	4227.2	4223.3	3.90	1.9
SMA	IFSTTAR	M3-2-7	d	19-7-2016	0	24.00	0	4158.5	4155.7	2.80	1.8
SMA	IFSTTAR	M3-2-12	b	21-7-2016	0	24.00	0	4216.5	4211.5	5.00	2.2
SMA	IFSTTAR	M3-2-15	c	25-7-2016	0	24.00	0	4222.8	4216.6	6.20	2.5
SMA	IFSTTAR	M3-3-3	d	25-7-2016	0	24.00	0	4183.5	4181.3	2.20	1.2
SMA	IFSTTAR	M3-3-12	a	28-7-2016	0	22.50	0	4129.1	4125.5	3.60	1.9
SMA	IFSTTAR	M3-3-13	b	1-8-2016	0	22.50	0	4194.1	4192	2.10	1.4
SMA	IFSTTAR	M3-3-14	c	2-8-2016	0	22.50	0	4074.5	4072.4	2.10	1.4
SMA	Heijmans	M3-1-6	a	6-9-2016	0	20.38	0	19174	19144	30.00	0.9
SMA	Heijmans	M3-1-9	a	7-9-2016	0	19.86	0	19063	19030	33.00	1.0
SMA	Heijmans	M3-1-11	a	9-9-2016	0	19.88	0	19460	19432	28.00	0.7
SMA	Heijmans	M3-1-17	a	15-9-2016	0	20.25	0	19296	19273	23.00	0.9
SMA	Heijmans	M3-2-2	a	16-9-2016	0	19.88	0	19168	19137	31.00	0.9
SMA	Heijmans	M3-2-8	a	20-9-2016	0	20.38	0	19091	19058	33.00	0.7
SMA	Heijmans	M3-2-11	a	21-9-2016	0	20.25	0	19410	19374	36.00	0.9
SMA	Heijmans	M3-2-20	a	22-9-2016	0	19.88	0	18978	18937	41.00	1.2
SMA	Heijmans	M3-3-6	a	26-9-2016	0	19.50	0	19457	19417	40.00	1.7

SMA	Heijmans	M3-3-7	a	27-9-2016	0	20.00	0	19413	19376	37.00	2.1
SMA	Heijmans	M3-3-8	a	28-9-2016	0	20.25	0	19335	19295	40.00	1.8
SMA	Heijmans	M3-3-15	a	29-9-2016	0	20.13	0	19454	19418	36.00	1.3
SMA	ISAC	M3-1-1	a	24-6-2016	0	29.13	0	17834	17817	17.00	
SMA	ISAC	M3-1-16	a	30-6-2016	0	26.13	0	17797	17784	13.00	
SMA	ISAC	M3-1-18	a	4-7-2016	1	26.63	0	17803	17790	13.00	
SMA	ISAC	M3-1-19	a	4-7-2016	2	28.25	0	17283	17268	15.00	
SMA	ISAC	M3-2-9	a	12-7-2016	0	28.63	0	18107	18093	14.00	
SMA	ISAC	M3-2-10	a	18-7-2016	1	27.50	0	17874	17865	9.00	
SMA	ISAC	M3-2-18	a	18-7-2016	2	29.25	0	17624	17613	11.00	
SMA	ISAC	M3-2-19	a	29-7-2016	0	30.38	0	17917	17907	10.00	
SMA	ISAC	M3-3-1	a	19-7-2016	1	28.88	0	17765	17753	12.00	
SMA	ISAC	M3-3-5	a	19-7-2016	2	30.25	0	17701	17691	10.00	
SMA	ISAC	M3-3-17	a	27-7-2016	0	29.88	0	17839	17822	17.00	
SMA	ISAC	M3-3-20	a	29-7-2016	0	32.88	0	17929	17916	13.00	
SMA	BAM	M3-1-10	a	1-7-2016	0	22.53	0	28298	28268	30.00	33.0
SMA	BAM	M3-1-13	a	5-7-2016	0	22.53	0	28148	28131	17.00	20.0
SMA	BAM	M3-1-15	a	6-7-2016	0	22.35	0	28829	28805	24.00	23.0
SMA	BAM	M3-1-20	a	10-8-2016	0	21.88	0	28113	28094	19.00	17.0
SMA	BAM	M3-2-3	a	8-7-2016	0	22.63	0	28725	28694	31.00	24.0
SMA	BAM	M3-2-4	a	13-7-2016	0	22.70	0	28610	28565	45.00	46.0
SMA	BAM	M3-2-14	a	22-7-2016	0	23.05	0	28686	28637	49.00	47.0
SMA	BAM	M3-2-16	a	21-7-2016	0	23.08	0	28221	28191	30.00	27.0
SMA	BAM	M3-3-2	a	27-7-2016	0	22.95	0	25261	25246	15.00	17.0
SMA	BAM	M3-3-10	a	29-7-2016	0	22.98	0	25524	25501	23.00	22.0
SMA	BAM	M3-3-11	a	2-8-2016	0	22.73	0	28320	28307	13.00	12.0
SMA	BAM	M3-3-19	a	8-8-2016	0	22.70	0	25381	25362	19.00	19.0

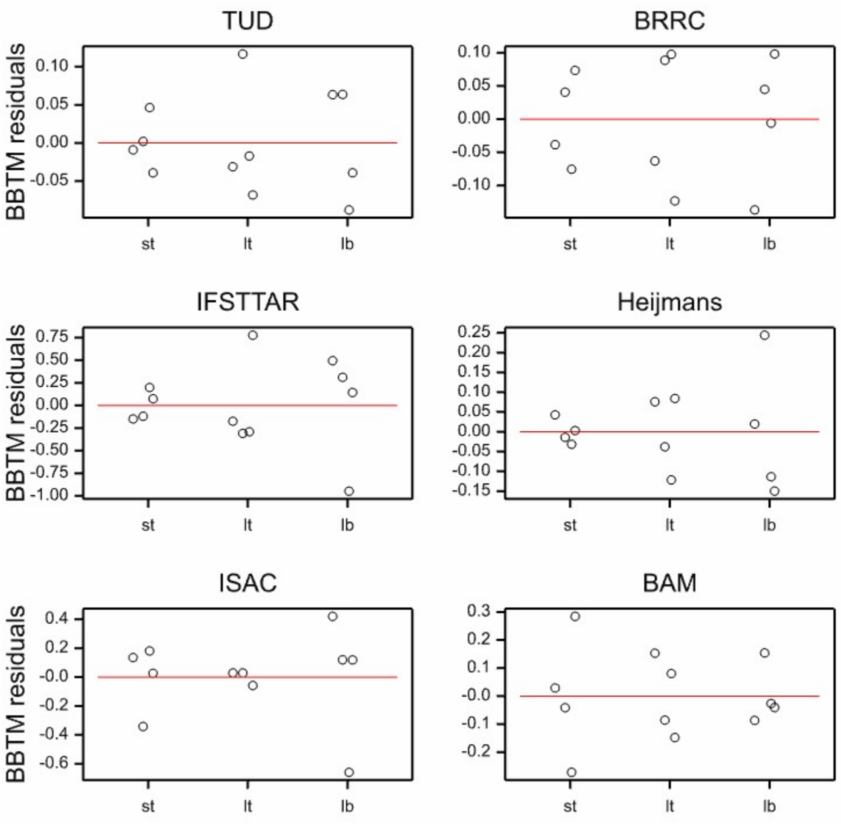
Appendix 2: Outlying weights

The reports of the various laboratories as well as our own inspection of the datasheets sent in by these laboratories identified six possible outlying weight losses in the data, as follows:

1. PA, TUD: weighing of M1-3-11 stopped early because the slab broke down.
2. PA, Heijmans: for slab M1-1-9, there was a problem with the shaft.
3. BBTM, IFSTTAR: for slab M2-2-11(b), measurements stopped after 7,000 cycles, as opposed to 10,000 cycles for the other three slabs of the same quality.
4. BBTM, IFSTTAR: the weight loss of slab M2-3-13(c) was more than 20 times smaller than the weight losses for the other three slabs of the same quality.
5. BBTM, ISAC: for slab M2-2-18, there was material loss at the edges.
6. BBTM, ISAC: for slab M2-3-17, the mass loss was 6 or 12 times smaller than the losses for the other three slabs.

Cases 1, 2 and 5 involve deviations from the measurement protocol. Results were removed from further analysis without further consideration. Cases 3, 4 and 6 were statistical outliers, identified by inspection of the residual plot below (a residual is a log(weight loss) minus the average of the log(weight loss) of the slabs in the same lab/quality set).

Statistical testing confirmed that the three weight losses were indeed outliers (P < 0.001 for both IFSTTAR outliers and P = 0.004 for the ISAC outlier). Inspection of the progress of deterioration of the slab during the testing confirmed the aberrant status of the slab.



Appendix 3: Means and standard deviations

The following tables show the geometric mean of the primary and secondary weight loss measures as well as the corresponding coefficients of variation (expressed as percentages of the mean).

Primary weight loss, geometric means

	Quality	s	lt	lb
Type	Lab			
PA	TUD	423.84	432.82	567.00
	BRRC	415.71	356.62	554.13
	IFSTTAR	1.09	1.69	1.26
	Heijmans	61.04	68.69	131.79
	ISAC	20.04	37.46	63.52
	BAM	42.50	67.99	96.84
BBTM	TUD	149.72	166.67	149.15
	BRRC	139.09	120.94	110.43
	IFSTTAR	9.89	2.81	34.83
	Heijmans	51.64	43.67	64.98
	ISAC	6.59	10.29	7.56
	BAM	18.70	18.28	32.96
SMA	TUD	7.51	5.49	5.82
	BRRC	4.74	5.25	4.73
	IFSTTAR	2.64	4.29	2.43
	Heijmans	28.26	35.05	38.21
	ISAC	14.41	10.85	12.76
	BAM	21.96	37.84	17.09

Secondary weight loss, geometric means

	Quality	s	lt	lb
Type	Lab			
PA	TUD	393.22	404.08	541.31
	BRRC	414.03	355.25	552.56
	IFSTTAR	0.85	0.90	0.59
	Heijmans	8.89	33.05	86.44
	ISAC	*	*	*

	BAM	45.25	69.70	88.59
BBTM	TUD	133.03	147.49	132.68
	BRRC	136.74	118.71	107.89
	IFSTTAR	*	*	*
	Heijmans	5.33	5.15	4.05
	ISAC	*	*	*
	BAM	18.09	18.05	32.24
SMA	TUD	4.39	3.01	3.36
	BRRC	4.38	4.09	3.51
	IFSTTAR	0.81	2.08	1.45
	Heijmans	0.87	0.91	1.70
	ISAC	*	*	*
	BAM	22.54	34.40	17.09

Primary weight loss, coefficient of variation

	Quality	s	lt	lb
Type	Lab			
PA	TUD	22.90	13.29	4.80
	BRRC	17.29	20.13	16.21
	IFSTTAR	53.76	44.07	60.29
	Heijmans	39.01	48.68	23.79
	ISAC	59.33	23.07	56.42
	BAM	17.80	30.32	23.63
BBTM	TUD	8.17	18.60	17.51
	BRRC	15.85	25.39	23.16
	IFSTTAR	37.94	16.74	40.42
	Heijmans	7.35	22.67	41.07
	ISAC	54.59	11.59	40.02
	BAM	52.74	32.34	24.38
SMA	TUD	31.25	48.66	29.00
	BRRC	32.79	31.15	26.31
	IFSTTAR	43.17	34.17	26.27
	Heijmans	15.29	12.11	5.41
	ISAC	12.93	18.86	22.05

	BAM	25.28	25.20	25.25

Secondary weight loss, coefficient of variation

	Quality	s	lt	lb
Type	Lab			
PA	TUD	21.81	13.99	5.56
	BRRC	17.21	19.68	16.47
	IFSTTAR	69.31	46.92	80.55
	Heijmans	47.87	39.18	25.73
	ISAC	*	*	*
	BAM	25.28	21.95	23.44
BBTM	TUD	6.92	20.95	20.43
	BRRC	16.56	25.69	22.48
	IFSTTAR	*	*	*
	Heijmans	22.15	29.08	18.92
	ISAC	*	*	*
	BAM	51.86	32.35	27.90
SMA	TUD	48.89	45.91	44.23
	BRRC	20.12	40.23	35.83
	IFSTTAR	24.76	14.84	19.26
	Heijmans	15.16	22.03	19.99
	ISAC	*	*	*
	BAM	28.26	35.12	25.82

Appendix 4: Detectable effect sizes

lab	asphalt	4 slabs	3 slabs	2 slabs
TUD	PA	1.4	1.5	1.9
TUD	BBTM	1.4	1.5	1.9
TUD	SMA	2.1	2.5	4.4
BRRC	PA	1.4	1.6	2.0
BRRC	BBTM	1.5	1.7	2.4
BRRC	SMA	1.8	2.1	3.3
IFSTTAR	PA	2.9	3.7	8.3
IFSTTAR	BBTM	2.0	2.3	3.9
IFSTTAR	SMA	2.0	2.4	4.1
Heijmans	PA	2.2	2.6	4.6
Heijmans	BBTM	1.7	2.0	3.0
Heijmans	SMA	1.3	1.3	1.6
ISAC	PA	2.7	3.4	7.1
ISAC	BBTM	2.3	2.8	5.3
ISAC	SMA	1.4	1.6	2.1
BAM	PA	1.6	1.8	2.6
BAM	BBTM	2.2	2.6	4.6
BAM	SMA	1.7	1.9	2.7