# Investigating the Reliability of Bioassays in Ecotoxicology – Addressing Questions of Reproducibility, Uncertainty and Interpretation

Faetsch, Sonja; Heise, Susanne; Karrasch, Maja; Kramer, Annette; Schlichting, Nicole


# Sullied Sediments

## Sediment Assessment and Clean Up Pilots in Inland Waterways in the North Sea Region

Many of the inland waterways in Europe are under threat due to the introduction of Watch List chemicals that are not currently regulated under the European Water Framework Directive. These chemicals enter our waterways as a result of our day-to-day activities and through industry, and many have been shown to be harmful to wildlife and the wider aquatic environment. Regardless of their source, these pollutants accumulate in the sediments in our rivers and canals over time.

Water regulators and managing authorities do not always know the levels, locations or impacts of these pollutants. Nor do they have the tools to assess sediments confidently and make informed environmental management decisions. To address these issues, the Sullied Sediment project partnership of scientific experts, regulators and water managers is developing and testing new tools that will enable stakeholders to better assess, treat and prevent contamination from these chemicals. This work is being carried out at selected sites in the Elbe, Humber and Scheldt river catchments.

The intention of the Sullied Sediments project is therefore to help regulators and water managers make better decisions with regard to the management, removal and disposal of sediments, thereby reducing economic costs to private and public sector organisations, and the impact of these pollutants on the environment.

The partnership is also working to reduce the extent of chemicals entering the water system by raising awareness about what we, as consumers, are releasing into the environment through the use of common drugs and household products. This includes the involvement of volunteers in a sediment sampling initiative across the North Sea Region, which will inform and empower them as water champions in their local communities.

The Sullied Sediments project has been co-funded by the European Regional Development Fund through the Interreg VB North Sea Region Programme with match funding from the 13 partners involved. The project partnership includes public, private, community and voluntary sector organisations based in the United Kingdom, Germany, Belgium and the Netherlands.

The project has been supported under the Interreg VB North Sea Region Programme's third priority, which is focused on a Sustainable North Sea Region, and is led by the University of Hull (UK).

**Website:** northsearegion.eu/sullied-sediments

**Blog:** sulliedsediments.wordpress.com

**Twitter:**@SulliedSediment

# Contents

# Executive Summary/Abstract

Few national regulations in Europe require ecotoxicological tests when assessing sediment quality or deciding about the fate of dredged material (den Besten et al. 2003; den Besten 2007; Heise et al. 2020). While scientists argue that ecotoxicological data will improve the environmental safety of dredged material management decisions compared to a solely on chemical data based framework, stakeholders claim that reproducibility, precision and accuracy of ecotoxicological data warrant a critical assessment before costly decisions are based on potentially unreliable information.

In order to assess the reliability of biotest results from testing natural sediments, several investigations were carried out in the course of the Interreg project Sullied Sediment:

1) Among the biotest data of the sediment sampling campaigns carried out in the scope of the Sullied Sediments project, those of two standardised bioassays, the algae growth inhibition assay (AGI, based on ISO 8692:2012, also termed as FWAT) and the luminescence bacteria test (LBT, based on ISO 11348-2:2008) conducted at the Hamburg University of Applied Sciences (HAW) and at Institut Dr. Nowak (IDN GmbH & Co. KG) were compared (Chapter 1). Results differed for most tests considerably. In the AGI, two different test species were used which are both recommended in the respective ISO standard. Toxicities indicated by *Raphidocelis subcapitata* and measured at HAW were in about 50 % of samples higher than those from *Desmodesmus subspicatus*, tested by IDN. LBT tests of the 2 labs also showed a trend towards higher toxicities measured at HAW than at IDN. Variations between the labs' procedures regarding elutriate preparations, time between sampling and test start, elutriate processing, but also different sensitivity of algae species were identified as potential causes, which are investigated in more detail in Chapter 3.

2) A round robin test was performed in October and November 2018 with six participating laboratories, using the AGI and the LBT as test systems. Outcomes of the test are presented in Chapter 2, while detailed statistical analyses are given in the annex. As far as we know, it represents the first round-robin test for both tests on natural sediments that included five to six independent laboratories. Instructions for sampling, sample handling and preparation of the test material (eluates and pore waters) were stipulated to minimise inconsistencies of results based on different treatments. The round robin test was supposed to answer the following questions:

(i) Do the standardized sampling and homogenization routines improve reproducibility of AGI and LBT results? (ii) How high is the variability of ecotoxicological endpoints (LID or pT-values)? (iii) Are AGI and LBT suitable to assess toxicity of natural sediments? And (iv) have test results any relevance for further analyses or interpretation of results?

Reproducibility of test results in this round robin test was poor, and the variability of results obtained from different laboratories was up to 200% for the AGI and up to 100% for the LBT. This indicates that the applied standardized sampling and homogenization routines and SOPs for the laboratories did not lead to a satisfying reproducibility of AGI and LBT results between different laboratories but might improve the repeatability of test results within the same laboratories as seen for the hidden duplicate samples from the LBT. Generally, the LBT seems to be less sensitive then the AGI and showed slightly better reproducibility than the AGI. The high number of invalid tests in single laboratories points to possible shortcomings with test performance. However, the data base of the present round robin test, especially the number of participating laboratories, is too poor to derive robust statistical results. Therefore, further studies to improve understanding of biotest responses and reliability of biotest results are required. According to our study, single biotest results from sediments that are achieved with the given kind of instructions (DIN standards and SOPs) should not be used as the sole indicators of a toxic potential of sediments because of the high interlaboratory variations. However, tendencies for differences in toxicity could be seen by integration of all test results. This should be taken into account when using data for sediment management decision.

3) In order to look for potential reasons for the interlaboratory variations, noticed in the Round Robin test and the results from the sampling surveys, another study was carried out to investigate the influence of sample handling and preparation of elutriates on test variance (Chapter 3). Two labs at HAW and Institut Dr. Nowak (IDN) tested elutriates

prepared from three sediment samples collected at the Elbe in Hamburg, applying the AGI and the LBT. The origin of observed variances of the test results was discussed. The time span between sampling and test start seemed to be very important. In case of the luminescent bacteria test the centrifugation of elutriates and their turbidity were identified as key factors influencing test results, showing the importance of more detailed instructions for test material preparation. In addition to procedural differences in the labs (miniaturization in 24 well plates versus performance in Erlenmeyer flasks), sensitivities of unicellular green algae species *Raphidocelis subcapitata* and *Desmodesmus subspicatus*, both eligible for use in guideline-based growth inhibition assays, seemed to differ significantly in some samples. These studies show that even though biotests, in principle, have the potential to provide an important additional line of evidence to sediment and dredged material assessment, small deviations from SOPs can lead to different outcomes.

The authors of this study conclude that:

- In order to improve reliability and comparability of ecotoxicological tests, the most critical methodological steps need to be identified, communicated and strictly controlled during test procedures.
- The interlaboratory LBT-comparison showed that the extent to which sample preparation and handling affects the outcome of the test varies with sediment, and can lead to substantially different outcomes in the bioassays. With synchronized and harmonized procedures, interlaboratory variations were between 5 and 18 percentage points and thus almost matched the in-house variability arising e.g. from different batches of luminescent bacteria.
- As it has been done with chemical analytical methods, interlaboratory comparisons of ecotoxicological testing of sediments should be carried out much more often in order to train staff and establish routines.
- A trend towards slightly higher toxicity was observed in the data from HAW compared to data from IDN. These can derive from different testing protocols (such as miniaturized set up versus set up in vials). Systematic differences in response ranges can be taken into account when interpreting the outcome of the bioassays and should be addressed in the assessment scheme for sediment and dredged material which is in development by the Sullied Sediment project.
- An assessment scheme should be developed on the bases of an integration of biotest results. Individual bioassays should not be assigned such far-reaching decision-making power that they determine management decisions. Instead, more importance should be placed on the outcome of biotest batteries and an integrated assessment of the different biotest responses, following a weight of evidence approach. This may result in a decision framework that should be environmentally safer while not being more costly.

# Introduction

The idea to test whether sediments and sediment elutriates have an effect on organisms and use this to inform on sediment quality goes back to the 1970s (e.g. (Anderson 1977; Swartz et al. 1979)). Giesy et al. (1988) explained the advantage of complementing chemical analysis of sediments by toxicity tests with the large number of potentially toxic substances which make the assessment of potential biotic impacts time consuming and costly. He stated that by focusing on chemical concentrations, neither bioavailability nor interaction of substances could be accounted for. Among the first, to incorporate toxicity tests of elutriates, porewater and/or sediments into management guidelines, was the US Army Corps of Engineers. The US Environmental Protection Agency developed a decision-making framework for the management of dredged material in 1985 that included biotesting of dredged material in addition to chemical evaluation if aquatic disposal was the preferred option (Lee and Peddicord 1988). With the proposal of a sediment quality triad approach, which combines ecotoxicological, chemical and community based data, Chapman (1990) formulated a frequently used weight of evidence strategy to determine pollution-induced degradation.

In Europe, ecotoxicological testing is less widely applied and incorporated in regulations compared to the US. In an overview in 2003, den Besten et al. (2003) showed the large variation between European countries with regard to ecotoxicological data in decision making frameworks which ranged from a solely chemical based assessment to biotests playing an important role in the decision support systems. During a workshop in 2018, data were gathered on how regulations have been changed over time, also with regard to the role of biological effect-based data, and there was no clear trend (Heise 2019). The Netherlands, that had relied on bioassays for their dredged material assessment in 2003, have now excluded them from their regulation. Italy, on the other side, increased the extent of biotesting marine dredged material significantly over the last years. Some countries still have little experience with ecotoxicological tests of freshwater sediments for regulatory purposes (e.g. England), while regulations in others take information from ecotoxicological data into account (e.g. Germany, for federal waterways). Whether biotests are, have been or are supposed to be anchored in national regulations, there is often quite a bit of skepticism among stakeholders when decisions are supposed to be based on biotest data. This paper specifically addresses the aspect of reproducibility, precision and accuracy of ecotoxicological data which, as stakeholders claim, warrant a critical assessment, in order to avoid basing costly decisions on unreliable information.

# Chapter 1: Comparison of HPA and HAW algae growth inhibition and bacterial luminescence test results based on the Sullied Sediments data base

By Sonja Faetsch & Susanne Heise (HAW), Nicole Schlichting (IDN)

## Introduction

In the course of Work Package (WP) 3 of the Sullied Sediments project six sampling campaigns at three river catchments in the North Sea region were conducted: Sediments were sampled between 2017 and 2019 in the seasons spring, summer and autumn. Three sites each were sampled in the Elbe catchment (Germany), the Humber catchment (UK) and the Scheldt River basin district (Belgium). To ensure uniform sampling procedures by all partners, taking samples and their homogenization in particular followed a Standard Operating Procedure (SOP) developed within WP 3 (Teuchies & WP 3 members, 2017). The sampling sites represented different pollution pressures, ranging from low to moderately and highly polluted sediments. The chemical analyses showed that the sediment samples contained pollutant mixtures that comprise heavy metals, organic compounds as well as emerging contaminants such as endocrine active substances at a range of concentrations (Richardson et al. in prep.).

Testing of German samples, taken at the Elbe estuary, was carried out at two different laboratories (HAW and IDN). Elutriates that contain the readily soluble fraction of pollutants in a sediment sample, imitating the resuspension of sediments under oxidising conditions, were ecotoxicologically assessed in two biotest systems: The algae growth inhibition (AGI) test and the bacteria luminescence test (LBT). These results will be shown and discussed in this chapter.

## Methods

### Sampling

Six sampling campaigns at the tidally influenced part of the River Elbe in Hamburg, the second largest city in Germany, were conducted in autumn 2017, in spring, summer and autumn 2018 as well as in spring and summer 2019. The three sampling sites were located upstream and downstream to the Port of Hamburg and within the Port, close to one of the largest wastewater treatment plants in Germany. The first site “Stover Strand” was located upstream of Hamburg representing historical pollution pressures from industrial and mining activities upstream during the time of the former German Democratic republic (GDR) and Czechoslovakia. The second site, “Köhlbrand”, was located in the area of the Hamburg harbour. It receives suspended sediments from both upstream and downstream. It was located close to the effluent of Hamburg’s largest wastewater treatment plant (WWTP) Köhlbrand/Dradenau, and was expected to be to some extent in the emission plume of the wastewater, reflecting the city’s urban industrial and human activities (pharmaceutical residues in particular). The third site “Wedel”, assessed the sediments pushed upstream into the Elbe Estuary with the flood stream mixed with the sediments from upstream areas. It was located at the sediment trap near Wedel that reduces the dredging amounts for maintaining fairway water depth in the port area and traps lesser-

contaminated fine sediments before they reach the port area where they would further mix with higher polluted sediments.

Sampling was carried out from aboard a vessel using a Van Veen grabber. Fresh sediments of the upper layer (≤ 15 cm) were sampled. Sampling was carried out according to the SOP developed in WP 3 (Teuchies and WP3 partners, 2017). Physico-chemical water and sediment parameters (pH, temperature, salinity, redox potential, oxygen concentration), water depth, weather conditions and tidal stage were documented. After thorough homogenization of the sediments for at least 5 minutes, subsamples were filled into glass jars, filled level to the top to avoid excess air. Samples were transported cooled to the test laboratories as quick as possible. Upon arrival at the laboratories samples were tested as soon as possible, but not later than two weeks after sampling to avoid decomposition and degradation of pollutants. Until biotesting, samples were stored in the refrigerator at 4-7°C. The two biotest performing laboratories in the work group Applied Aquatic Toxicology at the *Hamburg University of Applied Sciences* (HAW, referred to as Lab 1) and *Institut Dr. Nowak* (IDN, referred to as Lab 2) have extensive experience in performing the AGI and LBT.

## Elutriate Preparation

Elutriates for the AGI and LBT were produced at Lab 1 by shaking sediment mixed with ultra-pure water and using a volumetric method, following the American Society for Testing and Materials Guide E 1391 (ASTM 2000) und USEPA-U.S. Army Corps of Engineers (USEPA-USACOE 1998). The samples were mixed in a 1:4 (v/v) ratio of sediment to water and placed on an overhead shaker for 24 h, at a speed of ca. 7 rpm at room temperature (RT) in the dark. After centrifugation of the samples at 10.000 g for 20 min, the aqueous fraction (elutriate) was decanted and used for biotesting.

Lab 2 prepared elutriates based on their dry weight, according to the “Basisdokument zur Probenvorbereitung von limnischen Sedimenten für die Durchführung von limnischen Biotestverfahren im Auftrag der Hamburg Port Authority” (Floeter, 2009a) and the "BfG-Merkblatt "Ökotoxikologische Baggergutuntersuchung"” of the Federal Institute of Hydrology. Samples were mixed in a 1:4 (w/w) ratio, one part of dry weight of sediment and three parts of water. This water is the sum of pore water and synthetical freshwater, a dilution water named "Daphnienverdünnungswasser" stipulated in DIN 38412 T30. It contains several ions like sodium, magnesium, potassium and calcium. In the mixing container at least half the volume should be empty. If the water content was ≥ 75% the sample was eluted with its own water content. Samples were placed on an overhead shaker for 24 h, at ca. 7 rpm and RT. After centrifugation at 17.000 g for 20 min, the aqueous fraction was left open in a beaker for 30 min. Afterwards the parameters pH, conductivity and oxygen saturation were determined. If the oxygen concentration was ≤ 5 mg/l, samples were further aerated on a magnetic stirrer. Developing turbidity was removed by a second 10 min centrifugation step if necessary.

A series of dilutions was prepared from the elutriates. The initial elutriate (G1) was stepwise diluted with synthetical freshwater according to the applied guideline in 1:2 dilution steps until the inhibition was ≤ 20% compared to unexposed controls (negative control). The dilution grade G corresponds to the reciprocal of the volume fraction of the initial elutriate. If an elutriate was diluted 8 times (1:8) for instance, the dilution grade was G8. Due to the different protocols used for elutriate preparation, the initial elutriates and the dilution series prepared from them, corresponded to different percentages of the pore water fraction of the original sediment sample at the two laboratories. The percentages of pore water fractions in % of the elutriates were calculated for each dilution grade (Annex 1).

## Biotesting

### Algae Growth Inhibition Test

The AGIs using *R. subcapitata* and *D. subspicatus* are conducted following DIN EN ISO 8692:2012 at Lab 1 and DIN 38 412-L33 (1991) at Lab 2. They are conducted in a climate cabinet with defined temperature, light intensity and mixing by orbital shaking over 72 hours. The effect of elutriates on algae growth are determined after 72 hours, measuring

the chlorophyll fluorescence as surrogate parameter for biomass using multimode reader (Tecan infinite200). The response (% inhibition) is evaluated as a function of growth in the dilution grades of the elutriate in comparison with the average growth of negative controls. The algae represent the trophic level of primary producers.

At Lab 1 the test is carried out with the unicellular freshwater green algae *Raphidocelis subcapitata* (SAG strain 61.81, formerly known as *Pseudokirchneriella subcapitata*) in a miniaturized form in 24 well plates. A detailed SOP developed within the work group is used. Lab 2 uses the green algae *Desmodesmus subspicatus* (SAG strain 86.81) as test organism. In addition to the DIN Norm it applies the detailed manual "Basisdokument zum Grünalgentest mit *Desmodesmus subspicatus* basierend auf DIN 38 412-L33 (1991) und DIN EN ISO 8692 (2005) für die Bewertung von Sedimenten/ Baggergut im Auftrag der Hamburg Port Authority, Stand 2009_03 " (Floeter, 2009b). The test is conducted in Erlenmeyer flasks.

Two related endpoints can be derived: The fluorescence per well after 72 hours of incubation (T3) as surrogate parameter for algal biomass, and the growth rate per day, derived from the phase of exponential growth.

Only valid tests meeting the following performance criteria as stated in the respective norms were considered.

### Luminescence Bacteria Test

The LBT is carried out using liquid-dried luminescent bacteria *Aliivibrio fischeri* (Lumistox LCK482, Hach) as test organism, following DIN EN ISO 11348-2:2008 at Lab 1 and ISO 11348-2:2009-05 at Lab 2. It represents the inhibitory effects of pollutants present in the elutriates on bacteria. Bioluminescence is inhibited, if the bacterial energy metabolism is adversely affected by e.g. harmful substances,. Measurements are performed after a 30 min contact interval with a luminescence detector. At IDN, the measuring device LUMIStox 300, able to perform an internal colour correction, is used which works with single cuvettes and temperature controlling blocks. At HAW, the test is carried out in 96 well plates. All materials and solutions are kept at 15 °C prior to being measured in the luminescence reader (Tecan infinite200).

Only valid tests meeting the following performance criteria were considered: The mean of the correction value after 30 min ($fk_{30}$) ranged between 0.6 and 1.3, the mean coefficient of variation for luminescent measurements in the control cultures after 30 min did not exceed 5% and the inhibition of the positive control (4.5 mg/l of reference substance 3,5-Dichlorphenol) ranged between 20-80%.

## Results

### Algae Growth Inhibition Test

Due to the different preparation of elutriates the tested dilutions grades contained different fractions of the initial pore water at labs 1 and 2. In order to be able to compare the measured effects, for each sample and dilution step and both labs, the percentage of the original pore water of the sample, that was present in the respective tested elutriate, was calculated, using the information of pore water content of the sediment and the respective SOPs for elutriate preparation. The resulting pore water fractions within the aqueous sample were calculated and used as a common scale basis. They were plotted against the respective growth inhibition to compare results between the two labs. Even though these are discrete data points, they are connected by lines in the figures for easier demonstration. Plots of campaigns 3 and 5 are shown, representing all occurring response types (Fig. 1 and 2). In about 50% of the tests, *R. subcapitata* seemed more sensitive under the given conditions to the tested elutriates than *D. subspicatus*, e.g. sites 1 and 2 of campaign 3 (Fig 1 and 2). In ca. 50% of the tests sensitivity of the two species was rather similar, e.g. sites 2 and 3 of campaign 5 (Fig 2). Only in one case, site 3 of sampling campaign 3, *D. subspicatus* seemed more sensitive than *R. subcapitata*. Sensitivities generally converged in the lower pore water fractions. Most of the samples showed negative growth inhibition, especially in the lower dilution grades, which is probably caused by growth stimulating nutrients present in the samples.

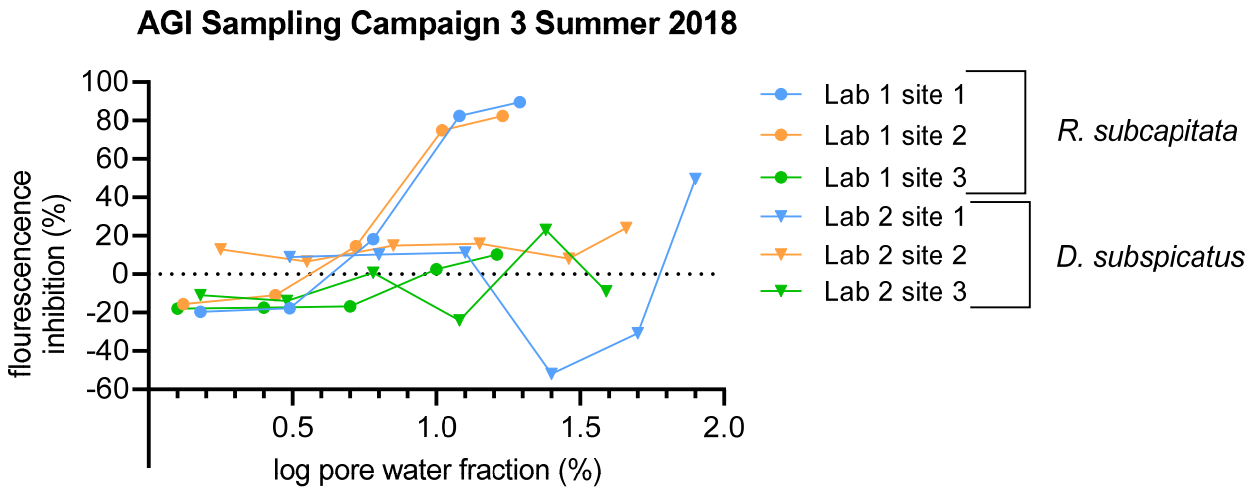



*Figure 1: Mean values of algae growth inhibition determined by fluorescence measurements of elutriates compared to unexposed control cultures. Results of algae growth inhibition test of Sullied Sediments sampling campaign number 3 in summer 2018. Circles: Lab 1 using, R. subcapitata. Triangles: Lab 2, using D. subspicatus. Blue: Site 1, Orange: Site 2, Green Site 3.*

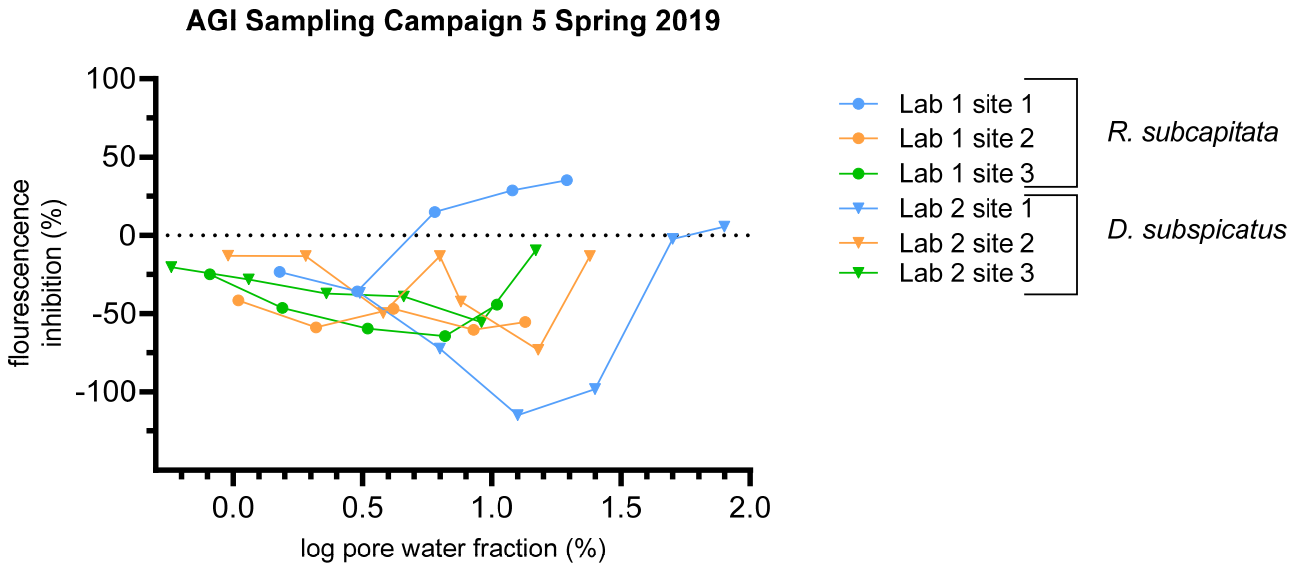



*Figure 2: Mean values of algae growth inhibition determined by fluorescence measurements of elutriates compared to unexposed control cultures. Results of algae growth inhibition test of Sullied Sediments sampling campaign number 5 in spring 2019. Circles: Lab 1 using, R. subcapitata. Triangles: Lab 2, using D. subspicatus. Blue: Site 1, Orange: Site 2, Green Site 3.*

## Luminescence Bacteria Test

Due to the different preparation of elutriates the tested dilution grades contained different fractions of the initial pore water at labs 1 and 2. The different pore water fractions were plotted against the respective luminescence inhibition to compare results between the two labs. Results of the LBT differed between labs. For the majority, 83% of samples, LBT at Lab 1 toxicities were considerably higher than at Lab 2, e.g. sites 1-3 of campaign 2 (Fig. 3, with 40 to 60 % at a pore water fraction of 10 % at Lab 1 compared to below 20 % at Lab 2). In only one case, LBT at Lab 2 was more toxic than at Lab 1 (site 1 of campaign 1, graph not shown here). For two samples, 11% of all samples, toxicities at both labs were very similar (site 1 of campaign 5, Fig 4). Toxicities at Lab 1 mostly showed steep increases with increasing pore water fractions, while at Lab 2 slopes were flatter (e.g. sites 1-3 of campaign 2, Fig 3).

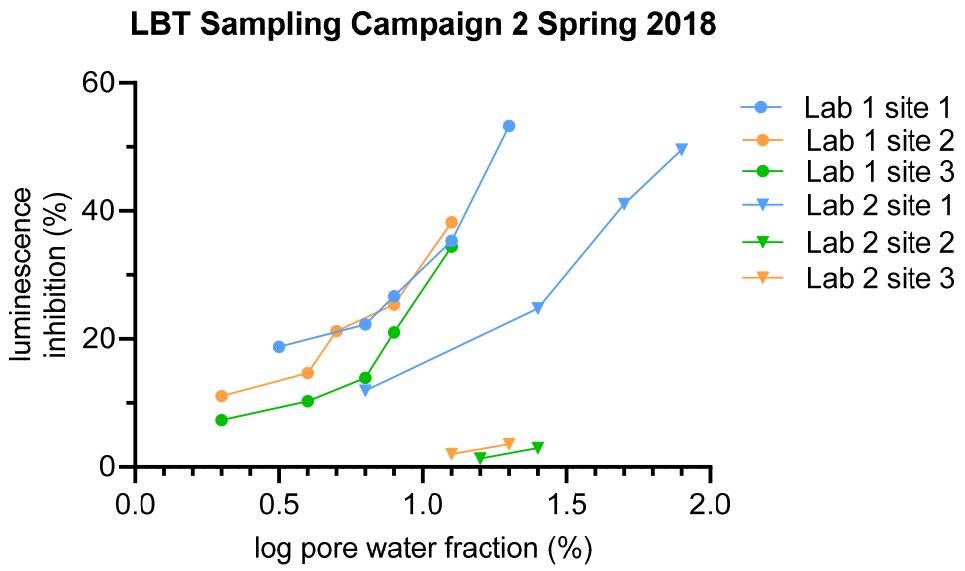



*Figure 3: Mean values of bacterial luminescence inhibition in elutriates compared to unexposed control cultures. Results of Luminescence bacteria test of Sullied Sediments sampling campaign number 1 in spring 2018. Circles: Lab 1. Triangles: Lab 2. Blue: Site 1, Orange: Site 2, Green Site 3.*

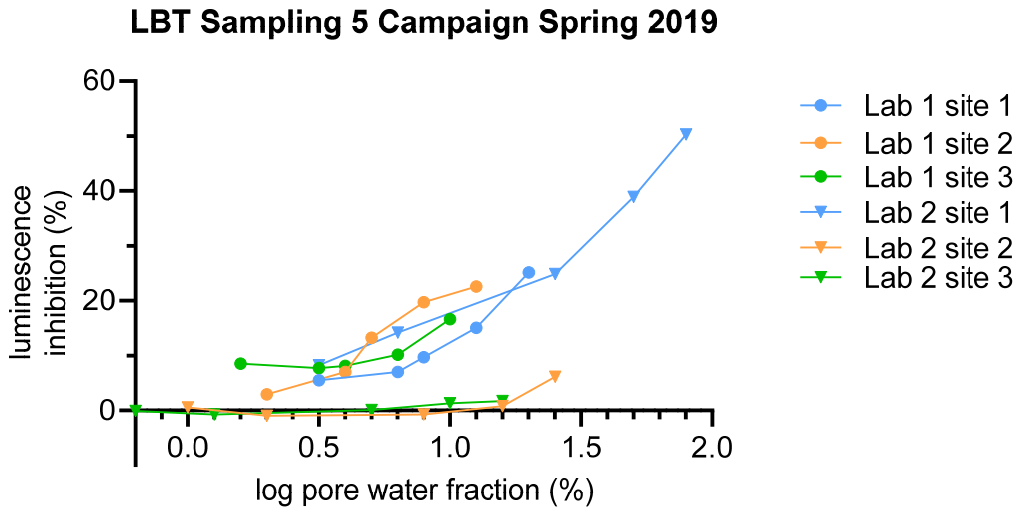



*Figure 4: Mean values of bacterial luminescence inhibition in elutriates compared to unexposed control cultures. Results of Luminescence bacteria test of Sullied Sediments sampling campaign number 1 in spring 2019. Circles: Lab 1. Triangles: Lab 2. Blue: Site 1, Orange: Site 2, Green Site 3.*

# Discussion

The Sullied Sediments project provided the opportunity to compare AGI and LBT test results of six sampling campaigns at three sites of the Elbe in the Hamburg port area. Tests were run at two laboratories that followed in principle the same norms for biotests but used different preparatory steps and different SOPs. A first comparison showed quite different elutriate preparations, starting by different ratios of sediment and dilution water, the dilution water itself, over to conditions during elution and centrifugation. Though it's not predictable what kind of effect those differences have on the solubility of substances, the link between inhibition and pore water fraction is an attempt to compare the results of both laboratories. Those observations led to another comparison-series of AGI and LBT where further hypothesis should be taken into account, e.g. the time span for sample processing after elutriate preparation and different measuring devices (see Chapter 3).

Considering these identified variations in the methodical procedures of the two labs a number of hypotheses were formulated that could explain the different biotest results:

1. *The high variabilities in LBTs are caused by different time spans between arrival at the lab and processing of the samples.* The storage time until elutriate preparation could have a significant influence, altering the samples to different degrees.

2. *The different preparation of elutriates, especially the first elution step in which sediment is mixed for 24 hours with water volumes that differ between the two protocols, leads to a different desorption of substances from the sediment matrix and such therewith to non-comparable results.*

3. *Different species sensitivities of R. subcapitata and D. subspicatus are responsible for the differences of results between the two labs.*

Another important aspect is to look at details in test procedures at the two labs that may cause variations in results, e.g. the usage of different measuring devices. These factors might have a huge impact on test results. Harmonizing these details of sample preparation and test procedure could reduce variabilities to a meaningful extent and improve acceptance of biotests in environmental decision-making.

Therefore, another study was planned to test these hypotheses. This study was carried out by HAW and IDN on fresh sediment samples requiring a much more detailed SOP and the exchange of prepared elutriates. A further comparison-series of AGI and LBT were conducted to investigate the above-mentioned hypotheses to explain the variabilities between the two labs (chapter 3).

# Chapter 2 How to assess sediment quality by divergent ecotoxicological test results? - Results from an ecotoxicological round robin test

By Annette Kramer and Maja Karrasch, Hamburg Port Authority

## Introduction and background

The handling of dredged material in Germany is regulated by two major guidelines. For the relocation/deposition within inland waters the HABAB (2017), and for the relocation/deposition of dredged material in coastal waters the GÜBAK (2009) applies, but also agreements and action plans with local authorities exists in which handling with dredged material is defined. The requirements how to assess dredged material differ between these regulations. While chemical analyses that take guiding values into account are included in every regulation, the application of ecotoxicological test methods to assess sediment quality beyond quantitative measures is not mandatory in every case. However, for crucial dredging activities within the port of Hamburg ecotoxicological data need to be collected. The procedure follows the requirements indicated by the Federal Institute for Hydrology (BfG, 2011). It indicates distinct biotest sets for marine, limnic and brackish environments involving organisms that shall reflect different trophic levels. For limnic environments, the Freshwater Algae Test (AGI or FWAT), the Luminescence Bacteria Test (LBT) and the Daphnia Test are required, for marine and brackish environments the Marine Algae Test (MAT), the LBT and a sediment contact test with *Corophium volutator* are required. The LBT and AGI needs to be applied on eluates and pore waters. Assessment criteria for single tests is the lowest ineffective dilution (LID) that is necessary to have the proxy for toxicity (e.g. inhibition rates) reduced below 20% (defined within DIN EN ISO 11348-2 for the LBT and DIN 38412-33 for AGI). The LID is then translated to pT-values (pT: potential Toxicologiae) as negative binary logarithm of the LID while the highest pT value that reflects most toxicity is used to assign the sediment sample to toxicity classes that provide criteria for handling of dredged material (Table 1).

*Table 1: pT values and toxicity classes derived by biotest batteries (modified from BFG, 2011).*

| Highest ineffective dilution | Dilution factor | pT-max value | Toxicity class | | Categories for management decisions |
|---|---|---|---|---|---|
| Original | $2^{0}$ | 0 | 0 | No toxicity measurable | Material not contaminated |
| 1:2 | $2^{-1}$ | 1 | I | Very low toxicity | Material only little contaminated |
| 1:4 | $2^{-2}$ | 2 | II | Low toxicity | |
| 1:8 | $2^{-3}$ | 3 | III | Moderate toxicity | Material critically contaminated |
| 1:16 | $2^{-4}$ | 4 | IV | Elevated toxicity | |
| 1:32 | $2^{-5}$ | 5 | V | High toxicity | Material is hazardously contaminated |
| ≤ (1:64) | ≤ $2^{-6}$ | ≥ 6 | VI | Very high toxicity | |

The current guideline for the disposal of dredged material in coastal waterways recommend the performance of an impact hypothesis in case of a toxicity level above class II (GÜBAK, 2009). Further, the deposition of dredged material of the Port of Hamburg in the North Sea is prohibited according to the current approval if dredged material shows a

significant higher toxicity than the biotest results of a reference data set from the same region. Since biotest results influence management decisions, reliable biotest methods are highly desired. They should both enable environmentally safe handling with dredged material but should not prevent a comprehensible scope of action for waterway maintenance. During the last 15 years, the Hamburg Port Authority (HPA) gained substantial experiences with the application of biotests on natural sediments. Results often showed that there was no clear relationship between pollution level of sediments according to the chemical parameters measured (generally, more than 60 contaminants) and the toxicity deduced by biotest results (AGI, LBT, daphnia test, and sediment contact tests with *Arthrobacter globiformis* and *Corophium volutator*) and that also no clear spatial pattern of sediment toxicity levels could be found. These could be due to components which are not included in the chemical parameter set as well as to synergistic or antagonistic effects of contaminant mixtures and additionally also to a changing bioavailability of compounds in the course of the year but also confounding factors could play a role. It also turned out, that identified toxicities were frequently unreproducible in inter-laboratory comparisons but also with hidden duplicate samples or repeated measurements within the same laboratory. Therefore, substantial effort was put on the optimisation of standard operation protocols (SOP) on how to handle samples and how to conduct the different tests on natural sediments (Floeter, 2009 a,b). Even with these detailed SOP, difficulties continued which is documented by results of two separated campaigns where the same sediment samples were sent to two different laboratories for running the AGI test. The results showed, that 65% of the data in the first campaign in 2012 and 40 % of the data in the second campaign in 2015 were not reproducible and showing a difference of at least two dilution stages/pT-values (Annex 2) which is not tolerable for management decisions.

Based on the outcome of these tests, the discussions concentrated on the homogenisation processes during sampling and sediment handling in general that might have led to poor reproducibility of biotest results. For addressing these problems and getting further insights into challenges and potentials of ecotoxicological methods the project Sullied Sediments provided excellent conditions. Within work package 3 (Sediment Assessment) of this project, a SOP (Teuchies & WP 3 members, 2017) was developed to ensure synchronised sampling campaigns in three catchments (Elbe, Humber Scheldt) which also involves specifications on how and for how long the samples needs to be homogenised to make sure each laboratory analyses the same sediments.

Here we present results of a new and more extensive round robin test on the AGI and LBT that was conducted within the scope of Sullied Sediments. As far as we know, it represents the first round-robin test for both tests on natural sediments that included five to six independent laboratories. Test results should help answering the following questions: (i) Do the standardised sampling and homogenisation routines improve reproducibility of AGI and LBT results? (ii) How high is the variability of ecotoxicological endpoints (LID or pT-values)? (iii) Are AGI and LBT suitable to assess toxicity of natural sediments? And (iv) Have test results any relevance for further analyses or interpretation of results?

## Material and Methods

The round-robin test for LBT and AGI was conducted in October and November 2018 on seven natural sediments. The sediments were collected in the part of the river Elbe that is influenced by the tides (samples 1-4), in the Humber catchment (sample 5), in the Nature Reserve *Biosphärenreservat Mittelelbe* (sample 6*)* and in the Scheldt catchment (sample 7) between the 4$^{th}$ and 18$^{th}$ October 2018.

Sediment 1 represents a site in the Lower Elbe upstream to the Hamburg Port (federal state Lower Saxony). It is located close downstream to the weir Geesthacht which cuts up tide stream for further upstream areas. The site is influenced by suspended sediments coming from upstream that could be loaded with contaminants originating from sites in the Middle Elbe still bearing legacies of industrial and mining activities before 1990. Sediment 2 was sampled in the Port of Hamburg (federal state Hamburg). The site is regularly dredged and is located close to the effluent of a wastewater treatment plant. The site is influenced by the tidal current and receives suspended sediments from downstream and upstream areas. Sediment 3 was taken from the main river outside of the federal state of Hamburg at a sediment trap near Wedel, where fine material from downstream is caught. The site is regularly dredged by the WSV as federal authority. Sediment 4 was sampled within the Port of Hamburg at a site that is mainly influenced by sediments coming from the Middle Elbe still bearing legacies of industrial and mining activities before 1990. Sediment 5 was collected at the banks of the river Aire in the UK downstream to a wastewater treatment plant. The site is influenced by former industrial and mining activities. Sediment 6 was collected in a shallow lake in the nature reserve *Biosphärenreservat Mittelelbe* (federal state Saxony-Anhalt). It originated from 6m under the sediment surface from Holocene layers. Sediment 7 represents a site in Belgium downstream from a wastewater treatment plant in the Bovenscheldt.

Chemical analyses of the sediments were conducted by sullied sediments partners on all sediments that were regularly sampled within the project (sample 1-3, sample 5+7). Samples 4 and 6 were analysed in an external laboratory.

Samples 1, 2 and 3 were selected as hidden duplicate samples. Therefore, each lab received 10 samples. Samples from the UK and Belgium were sent to HPA where they were unitised as subsamples for the round robin test. All samples were homogenised using a stainless-steel mixer associated with a drilling machine, packed in 2.5l food safe plastic buckets and send out in cold boxes with thermal packs to the different laboratories via mail.

Selected laboratories were all based in Germany to keep time for transportation and mailing of samples short as samples for ecotoxicological analyses need to be analysed within two to three weeks after sampling. Participating laboratories are listed in alphabetical order in Table 2 that is not related to lab codes used in the assessment. The LBT was conducted in six laboratories, the AGI was conducted in five laboratories. All selected laboratories have a long expertise for biotest testing, only Laboratory 6 applied the LBT for the first time.

*Table 2: Participating laboratories in alphabetical order (from Ratte 2020)*

| **Laboratory** |
|---|
| ECT Oekotoxikologie GmbH |
| Eurofins Umwelt Ost GmbH |
| HAW Hamburg University of Applied Sciences (working group Applied Aquatic Toxicology) |
| Hydrotox GmbH |
| Institut Dr. Nowak GmbH & Co. KG |
| Noack Laboratorien GmbH |

For the laboratory work, SOPs to prepare pore waters and eluates (Floeter, 2009a) and a detailed manual that is based on DIN 38412-32 and DIN EN ISO 8692 how to conduct the AGI test (Floeter, 2009b) with the algae *Desmodesmus*

*subspicatus* were provided for all participating laboratories. The LBT was run following the DIN EN ISO 11348-2 and the advices from the Federal Institute for Hydrology (BfG, 2011).

Laboratories were asked to follow the protocol given by the round robin test organiser and to deliver test results in a table that includes all test parameters. However, deviation occurred in single labs and were considered for statistical analyses to make data better comparable. The following exceptions were adjusted as already described in Ratte (2020, Annex 3): **(i)** Some laboratories did not test the prescribed dilutions, but fewer ones or other ones. Generally, dilution series for LBT should have been proceeded until inhibition is below 20%. This was not done by laboratory 3 in every single test for sediment 1, 4, 7. Further, laboratory 5 did not test the inhibition values for the D1. Additionally, in the AGI, the critical inhibition of 20% in some tests was still exceeded in the highest dilution (D32). Consequently, in these cases, the LID could only be judged as < x or > x. To enable statistical evaluation, these results were set to x. **(ii)** For the LBT, laboratory 3 did not apply the colour correction for the inhibition values that should have been used for the summary table. Therefore, the corrections were calculated by the test organiser for each dilution stage based on the results from laboratory 1 who provided both colour corrected values and not corrected values. The inhibition values were thus slightly lower for laboratory 3 than it was reported by the lab. Laboratory 3 also applied a miniaturised LBT with smaller sample volumes. Every laboratory used DIN EN ISO 11348-2 conform bacteria while laboratory 4 used a different source of bacteria than the other laboratories. Further descriptions of the data base are given in Annex 3, chapter 2.

Additional to test parameter, characteristics of pore waters and eluate like nutrient and oxygen content, pH-values, salinity and conductivity needed to be recorded by participating laboratories. Test results (percentages in D1 and pT-values) were correlated to these parameters to find relationships that might explain contradictive test results.

Validity of test results was checked by the participating laboratories and re-checked by the round robin test organiser. For the present report all results were considered (further details see Ratte 2020, Annex 3). The statistical analyses were run by Ratte (2020) on the dilution stages (LID) and the inhibition rates in percentages in dilution stage 1 (D1) separately for pore water and eluates. Analyses were done twice: on all data and on only valid data. For each sediment the geometric mean and the coefficient of variation (CV) as percentages were calculated following the procedure given in Ratte (2020, Annex 3: chapter 2.3). The CV% serves as measure of the reproducibility for the test results. Additionally, the normalised width (NW) was used to better assess reproducibility for the inhibition percentages of D1. Detailed procedure of data processing is given in Annex 3.

For the present report, an additional consideration of each sample is done on basis of official sediment assessment by means of biotest data (BfG, 2011), where usage of maximal toxicities from single tests within the whole biotest battery to assess sediments is mandatory.

# Results

## Sediment and test material

Sediments differ concerning their physical and chemical properties. An overview of the results of the analyses is given in Table 3. Sediments 4 and 6 show a fine fraction (<63µm) share of more than 90%, while sample 5 shows only a fine fraction of 4,5%. Sample 1, 2, 3 and 7 have shares between 43 and 72%.

*Table 3: Results of physical-chemical analyses run by SOCOTEC (metals) and Institut Dr. Nowak (Organic substances, grain size, dry weight) as Sullied Sediment project partners. Samples 4 and 6 were analysed by an external laboratory. Only those parameters are shown that were analysed in all seven sediments. Maximal values of data set are printed in bold. Negative values represent detection limits.*

| Sample | | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | Sample 6* | Sample 7 |
|---|---|---|---|---|---|---|---|---|
| SuSe Sample ID | Unit | DE_4.1 | DE_4.2 | DE_4.3 | ReiV | UK_4.2 | Reference | BE_4.1 |
| dry weight (%) | % | 41.9 | 39.9 | 59.4 | 18.7 | 70.8 | 17 | 52.1 |
| > 2 mm | % | 0.80 | 0.80 | 0.90 | <0.1 | 17.10 | 0 | 0.30 |
| 1000-2000 µm | % | 0.60 | 0.30 | 0.50 | 0.2 | 3.70 | 0.5 | 0.40 |
| 600-1000µm | % | 1.90 | 0.30 | 1.60 | 0.4 | 6.70 | 1.1 | 0.40 |
| 200-600 µm | % | 14 | 2 | 7 | 0.4 | 52 | 1.6 | 4.40 |
| 100-200 µm | % | 19 | 7 | 16 | 1.1 | 14 | 1.1 | 24 |
| 63-100 µm | % | 8.60 | 18 | 30.90 | 3.3 | 2 | 1.6 | 16 |
| 20-63 µm | % | 19 | 29 | 14 | 23.4 | 1 | 7 | 23 |
| < 20µm | % | 36 | 43 | 30 | 71.2 | 3 | 87 | 31 |
| < 100 µm | % | 63 | 90 | 75 | 98 | 7 | 96 | 70 |
| < 63 µm | % | 55 | 72 | 44 | 95 | 5 | 94 | 54 |
| As | mg/kg DW | **20** | 16 | 13 | 27 | 9.8 | 3.6 | 8.1 |
| Cd | mg/kg DW | **3.30** | 0.82 | 0.55 | 2.80 | 1.00 | -0.1 | 2.50 |
| Cr | mg/kg DW | 64 | 68 | 48 | 44 | 75 | 3.1 | **140** |
| Cu | mg/kg DW | 30 | 13 | **77** | 58 | 69 | 16 | 48 |
| Hg | mg/kg DW | 0.58 | 0.30 | 0.20 | **0.71** | 0.09 | 0.02 | 0.20 |
| Ni | mg/kg DW | **35** | 32 | 22 | 30 | 30 | 4.7 | 27 |
| Pb | mg/kg DW | 34 | 52 | 41 | 57 | **87** | 1.5 | 64 |
| Zn | mg/kg DW | 310 | 170 | 440 | 490 | 210 | 19 | **560** |
| Monobutyltin cation | µg/kg DW | **25** | 23 | 16 | **25** | 15 | -1.0 | 16 |
| Dibutyltin cation | µg/kg DW | 15 | 9.0 | 6.0 | 16 | 4.0 | -1.0 | **19** |
| Tributyltin cation | µg/kg DW | 18 | **26** | 16. | 23 | 2.0 | -1.0 | 5 |
| Tetrabutyltin | µg/kg DW | **17** | 7.0 | 4.0 | 7.3 | -1.0 | -1.0 | -1.0 |
| Tricyclohexyltin cation | µg/kg DW | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 |
| Triphenyltin cation | µg/kg DW | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | -1.0 | 2.0 |
| Monooctyltin cation | µg/kg DW | 4.0 | 1.0 | -1.0 | 2.9 | 1.0 | -1.0 | **6.0** |
| Dioctyltin cation | µg/kg DW | 5.0 | 1.0 | -1.0 | 1.7 | 1.0 | -1.0 | **9.0** |
| PCB28 | µg/kg DW | 1.20 | 0.27 | 0.18 | 1.0 | 0.68 | -0.10 | **1.30** |
| PCB52 | µg/kg DW | 0.96 | 0.20 | 0.15 | 1.2 | 0.70 | -0.10 | **2.10** |
| PCB101 | µg/kg DW | 2.20 | 0.63 | 0.46 | 2.2 | 0.79 | -0.10 | **4.40** |
| PCB118 | µg/kg DW | 1.30 | 0.42 | 0.30 | 1.4 | 0.66 | -0.10 | **3.70** |
| PCB138 | µg/kg DW | 4.90 | 1.40 | 0.83 | 4.5 | 1.40 | -0.10 | **6.90** |
| PCB153 | µg/kg DW | 6.10 | 1.40 | 1.10 | 3.6 | 1.00 | -0.10 | **7.00** |
| PCB180 | µg/kg DW | 6.20 | 0.93 | 0.77 | 3.6 | 1.30 | -0.10 | **6.40** |
| PCB Sum | µg/kg DW | 22.86 | 5.25 | 3.79 | 16.1 | 6.53 | | **31.80** |
| aHCH | µg/kg DW | **3.90** | 0.27 | 0.17 | 0.44 | -0.10 | -0.05 | -0.10 |
| bHCH | µg/kg DW | **7.00** | 0.68 | 0.49 | 1.2 | 0.23 | -0.10 | 0.17 |
| yHCH | µg/kg DW | **1.10** | 0.11 | -0.10 | 0.18 | 0.22 | -0.05 | -0.10 |
| o-p DDD | µg/kg DW | **22.00** | 2.60 | 1.70 | 5.9 | 0.69 | -0.10 | 1.60 |
| p-p DDD | µg/kg DW | **54.00** | 6.40 | 4.60 | 12 | 3.10 | 0.11 | 1.70 |
| o-p DDE | µg/kg DW | **0.65** | -0.10 | -0.10 | 0.49 | -0.10 | -0.10 | -0.10 |
| p-p DDE | µg/kg DW | **7.70** | 1.50 | 1.00 | 5.3 | 0.69 | -0.10 | 2.00 |
| o-p DDT | µg/kg DW | **3.10** | 0.13 | -0.10 | 1.7 | -0.10 | -0.10 | -0.10 |
| p-p DDT | µg/kg DW | **24.00** | 0.53 | 0.34 | 3.7 | 0.25 | -0.10 | 0.30 |
| DDX | µg/kg DW | **111.45** | 11.26 | 7.84 | 29.09 | 4.93 | 0.61 | 5.80 |
| Hexachlorobenzene | µg/kg DW | 3.60 | 2.70 | 1.60 | **7.6** | 0.41 | -0.10 | 0.40 |

*Holocene sediment was analysed twice as first results were regarded as implausible (first analyses not shown).

The chemical analyses also reveal differences between the sediments. The sediment with the highest number (14) of maximal contaminant content among the here presented test set is sediment 1. It shows maximal values of arsenic, cadmium, nickel, MBT, TeBT, aHCH, bHCH, gHCH and of the DDT-compounds. Sample 7 shows second highest number

(13) of maximal contaminant contents with chrome, zinc, and all PCB congeners, DBT, Mono- and Dioctyltin as maximal contents. Sample 6 shows organic contaminants mostly under detection limits and lowest metal content of all samples.

Test material like pore waters and eluates also differs regarding the analysed nutrients and oxygen content before aeration between the laboratories (Table 4). Unfortunately, not all participating laboratories analysed the nutrient content of pore waters and eluates and test methods differ between laboratories. Therefore, results regarding nutrient contents have to be considered with care. But oxygen content and conductivity show that the produced and analysed test material differs in some cases.

Since nitrite and nitrate were measured mostly under the detection limit and every laboratory applied different detection limits, those parameters were not presented here.

Lowest mean oxygen values were obtained for sample 1 and 4 in pore waters and eluates. Mean oxygen values below 5mg/l were also determined for the pore waters for sediments 2 and 7. Highest ammonia values were detected in sample 1 and 4 and lowest in sediment 3 and 6. The conductivity is higher in each pore water than in the eluates with highest values found in sample 4 for both pore waters and eluates. Lowest values are seen in sample 5 and 6 (Table 4).

*Table 4: Results of test material analyses provided by the participating laboratories*

| Sample | | $O_2$ mg/l before aeration | | | Conductivity µs/cm | | | pH-Value | | | $NH_4$-N mg/l | | | total N mg/l | | | $PO_4$-P mg/l | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min | Max | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max | Mean | Min | Max | Mean |
| 1 | PW | 0.9 | 7.0 | 3.7 | 1819 | 2450 | 2186 | 6.9 | 7.5 | 7.2 | 47.0 | 102 | 64 | 10 | 73 | 46 | 0.02 | 2.3 | 0.7 |
| 1 | EL | 1.5 | 8.2 | 4.4 | 1775 | 2149 | 1940 | 7.0 | 8.0 | 7.2 | 53.9 | 108 | 70 | 12 | 76 | 51 | 0.01 | 1.3 | 0.2 |
| 2 | PW | 1.9 | 7.8 | 4.5 | 1858 | 2750 | 2274 | 7.1 | 7.9 | 7.5 | 10.6 | 16 | 14 | 10 | 22 | 17 | 0.03 | 1.2 | 0.4 |
| 2 | EL | 3.8 | 8.4 | 6.0 | 1374 | 2110 | 1671 | 7.1 | 7.8 | 7.3 | 5.0 | 13 | 9 | 16 | 59 | 27 | 0.02 | 0.9 | 0.2 |
| 3 | PW | 2.9 | 8.1 | 5.3 | 2070 | 2770 | 2384 | 7.5 | 7.8 | 7.6 | 6.0 | 10 | 9 | 10 | 24 | 14 | 0.03 | 1.3 | 0.4 |
| 3 | EL | 3.8 | 9.0 | 6.6 | 946 | 2430 | 1352 | 6.0 | 7.6 | 7.2 | 0.0 | 3 | 1 | 16 | 51 | 24 | 0.05 | 0.6 | 0.2 |
| 4 | PW | 0.2 | 7.5 | 2.1 | 2490 | 3080 | 2876 | 7.0 | 7.9 | 7.2 | 18.2 | 114 | 85 | 61 | 126 | 97 | 0.01 | 1.9 | 0.4 |
| 4 | EL | 2.9 | 7.4 | 5.0 | 2250 | 2830 | 2550 | 7.0 | 7.6 | 7.2 | 19.3 | 111 | 85 | 59 | 144 | 106 | 0.04 | 1.9 | 0.7 |
| 5 | PW | 4.3 | 8.4 | 6.1 | 994 | 1215 | 1113 | 6.9 | 8.0 | 7.5 | 7.1 | 11 | 9 | 7 | 12 | 10 | 0.04 | 3.0 | 0.7 |
| 5 | EL | 4.1 | 8.3 | 6.1 | 661 | 1493 | 1023 | 7.0 | 7.2 | 7.1 | 0.4 | 4 | 2 | 6 | 14 | 8 | 0.03 | 0.7 | 0.3 |
| 6 | PW | 4.7 | 8.1 | 6.9 | 1020 | 1320 | 1157 | 6.8 | 7.5 | 7.1 | 2.4 | 18 | 10 | 26 | 44 | 32 | 0.03 | 0.1 | 0.1 |
| 6 | EL | 4.8 | 9.2 | 7.2 | 882 | 1493 | 1228 | 7.0 | 7.4 | 7.2 | 1.0 | 29 | 7 | 28 | 45 | 35 | 0.08 | 0.7 | 0.4 |
| 7 | PW | 1.5 | 7.2 | 4.3 | 1930 | 2360 | 2177 | 7.0 | 7.5 | 7.2 | 17.4 | 60 | 45 | 48 | 72 | 60 | 0.03 | 8.5 | 4.0 |
| 7 | EL | 1.5 | 8.5 | 5.4 | 964 | 2280 | 1439 | 6.8 | 7.7 | 7.1 | 29.0 | 36 | 32 | 34 | 51 | 44 | 0.29 | 1.7 | 0.9 |

## Results obtained by the Luminescence Bacteria Test (LBT)

Based on the requirements given by the relevant DIN standards, 92.5% of the LBT data were valid (see Annex 2: Table 3 - Table 4). While the intra-laboratory variation for the LBT is low and most of the duplicates show similar results, test results of LBT vary between the different laboratories clearly (Figure 5).

For sediment 1 and sediment 4 results show a range between a pT-value of 0 and 4 signalling a toxicity between no toxicity at all and elevated toxicity (see Table 1). Little less variation is visible for the sediments 2, 3, 5 and 6 and pT-values between 0 and 2 (no toxicity at all and low toxicity) were deduced. Sediment 7 shows results between pT-vales of 0 and 3 (no toxicity at all and moderate toxicity. Systematic differences between results of PW or EL are not visible

in the data set. When comparing the inhibition rates obtained for each sediment in D1 for PW and EL similar offsets between results of different laboratories are obvious (Table 5). Sample 1 and 4 received highest inhibition rates but also negative values were obtained pointing to stimulating instead of inhibition effects of test material on the bacteria.

| LBT | | Lab 1 | Lab 2 | Lab 3 | Lab 4 | Lab 5 | Lab 6 | | Maximal toxicity needs to be considered for assessment of sediment (results LBT) |
|---|---|---|---|---|---|---|---|---|---|
| | | pT | pT | pT | pT | pT | pT | | |
| **Sample 1** | PW | 0 | 2 | 3 | 0 | 1 | 1 | | 4 |
| | EL | 2 | 2 | 4 | 0 | 4 | 0 | | |
| Duplicate | PW | 0 | 2 | 3 | 0 | 1 | 1 | | 4 |
| | EL | 2 | 2 | 4 | 0 | 3 | 1 | | |
| **Sample 2** | PW | 0 | 0 | 2 | 0 | 1 | 1 | | 2 |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | | |
| Duplicate | PW | 0 | 0 | 1 | 0 | 1 | 0 | | 1 |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | | |
| **Sample 3** | PW | 0 | 0 | 1 | 0 | 1 | 1 | | 1 |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | | |
| Duplicate | PW | 0 | 0 | 1 | 0 | 1 | 1 | | 1 |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | | |
| **Sample 4** | PW | 2 | 1 | 4 | 0 | 2 | 0 | | 4 |
| | EL | 2 | 1 | 3 | 0 | 1 | 1 | | |
| **Sample 5** | PW | 2 | 1 | 2 | 0 | 2 | 0 | | 2 |
| | EL | 0 | 0 | 2 | 0 | 1 | 0 | | |
| **Sample 6** | PW | 0 | 0 | 1 | 1 | 1 | 0 | | 1 |
| | EL | 0 | 0 | 1 | 0 | 1 | 0 | | |
| **Sample 7** | PW | 0 | 1 | 3 | 0 | 1 | 0 | | 3 |
| | EL | 0 | 0 | 1 | 0 | 1 | 0 | | |

*Figure 5: Results of LBT presented as pT-values for pore waters (PW) and eluates (EL) for each laboratory and sediment. Sediments 1-3 were introduced as hidden duplicate samples[1].*

*Table 5: Minimum and maximum inhibition rates % in D1 obtained by LBT in laboratories 1-4 and 6 for the sediments*. Values in brackets show minima and maxima of inhibition rates % when results from Lab 4 who used a different producer of bacteria were excluded.*

| Sediment | | Inhibition rate % Min | Inhibition rate % Max |
|---|---|---|---|
| Sample 1 | PW | -79 (15) | 59 |
| Sample 1 | EL | -38 (-6) | 87 |
| Sample 2 | PW | -70 (2) | 33 |
| Sample 2 | EL | -62 (-7) | 16 |
| Sample 3 | PW | -73 (5) | 31 |
| Sample 3 | EL | -69 (-7) | 16 |
| Sample 4 | PW | -73 (5) | 87 |
| Sample 4 | EL | -13 (33) | 67 |
| Sample 5 | PW | -24 (1) | 72 |
| Sample 5 | EL | -36 (-5) | 58 |
| Sample 6 | PW | -26 | 58 (41) |
| Sample 6 | EL | -49 (-6) | 46 |
| Sample 7 | PW | 11 | 63 |
| Sample 7 | EL | -50 (1) | 50 |

*Laboratory 5 did not measure inhibition rates in D1

When applying the valid criteria for sediment assessment (BfG, 2011) sediment 1, 4 and 7 should be treated as material that is critically contaminated. Sample 2, 3, 5 and 6 are classified as material that is only little contaminated (Figure 5)

[1] Laboratory 5 did not measure inhibition rates in D1

In the statistical analyses by Ratte (2020) of the round robin test mean values of the LID were calculated for each sediment. Therefore, the values differ from the official LID or pT-value classification. Highest mean values were obtained for sediment 1 for the eluates and for sample 4 for the pore waters and lowest mean values for sediment 2 and 3 for the eluates and sediment 6 for the pore waters (Figure 6).

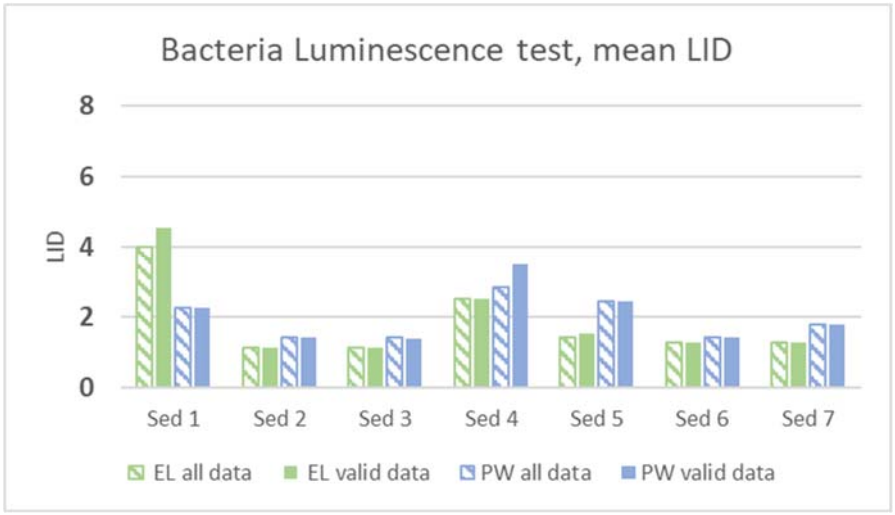



*Figure 6: Mean of LID obtained in the Bacteria Luminescence test. EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data, (taken from Ratte (2020))*

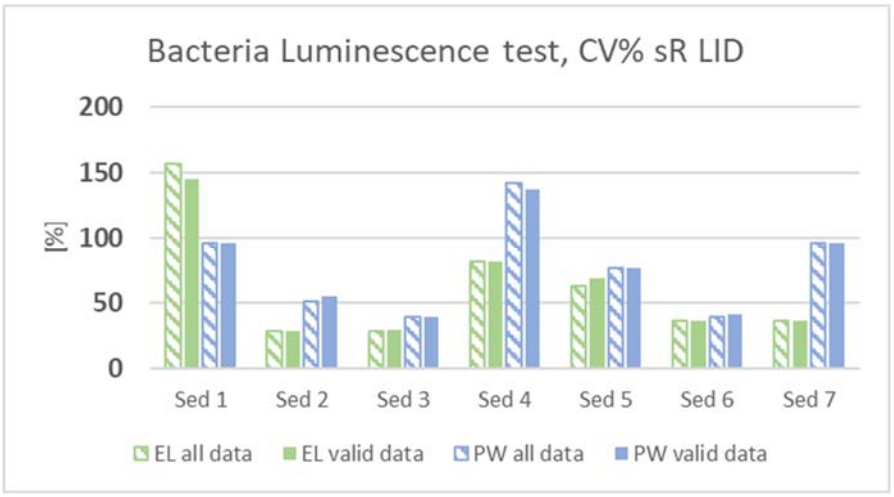



*Figure 7: CV% for reproducibility of LID obtained in the Bacteria Luminescence test. EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data (taken from Ratte (2020))*

The coefficient of variation revealed a variation around or above 100% for samples 1, 4, 5 and 7. For samples 2, 3 and 6 variations were around 50% or lower (Figure 7).

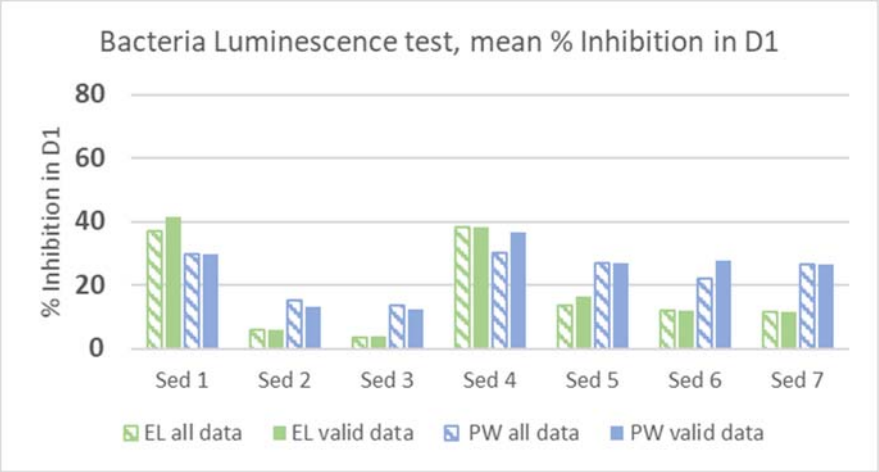



*Figure 8: Mean of inhibition rate (%) in D1 obtained in the Bacteria Luminescence test. EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data (taken from Ratte (2020)).*

The mean inhibition rates for D1 were additionally calculated. In doing so, negative inhibition rates were set to 0 (see Ratte, 2020, Annex 3). Mean inhibition values above 20% were obtained for sample 1 and 4 for the eluates and for sediment 1, 4, 5, 6 and 7 for the pore waters (Figure 8). Except for sample 5 in the eluates and sample 1 for the pore waters the normalised width (NW) shows values above or equal 2 meaning "that the true mean will be within a range of twice the calculated mean" (Ratte 2020).

The intra-laboratory variability for the LBT was assessed by analysing duplicates for sediments 1-3. In all 36 duplicate samples of sediment 1, 2 and 3, identical LIDs (32) or LIDs which differ not more than one dilution step (4) were obtained (Ratte, 2020). Due to statistical reasons no CV could be obtained for the LID data, but on the inhibition rates. CV% for the D1 show that CV% was between 35% and 57% for EL samples, for PW samples CV% was between 6% and 14% (Ratte, 2020).

## Results obtained by the Fresh Water Algae Test (AGI)

Based on the requirements given by the relevant DIN standards, 76% of the AGI data were valid (see Annex 2: Table 2 and Table 4).

For sediment 1, 2, 3 and 4 AGI results show a range between a pT-value of 0 and 5 signalling a toxicity between no toxicity at all and high toxicity (Figure 9, see also Table 1). Little less variation is visible for the sediments 5 and 7 and pT-values between 0 and 2 (no toxicity at all and low toxicity) were deduced. Sediment 6 shows results between pT-vales of 0 and 3 (no toxicity at all and moderate toxicity). Systematic differences between results of PW or EL samples are not visible in the data set. When comparing the inhibition rates obtained for each sediment in D1 for PW and EL samples, similar offsets between results of different laboratories are obvious by presenting differences between inhibition rates in one sample of up to 160% (EL, sample 2) (Table 6). Sample 1, 2, 3 and 4 received highest inhibition rates above or equal 80% but also negative values were obtained for each sample pointing to promotion instead of inhibition effects.

The intra- and inter-laboratory variation for the AGI is relatively high as visible in Figure 9.

| FWAT | | Lab 1 | Lab 2 | Lab 4 | Lab 5 | Lab 6 | | Maximal Toxicity needs to be considered for assessment of sediment |
|---|---|---|---|---|---|---|---|---|
| | | pT | pT | pT | pT | pT | | |
| **Sample 1** | PW | 1 | 2 | 4 | 1 | 0 | | 4 |
| | EL | 2 | 1 | 0 | 1 | 0 | | |
| Duplicate | PW | 1 | 5 | 4 | 1 | 0 | | 5 |
| | EL | 1 | 2 | 0 | 2 | 0 | | |
| **Sample 2** | PW | 0 | 0 | 4 | 0 | 0 | | 4 |
| | EL | 1 | 2 | 4 | 3 | 3 | | |
| Duplicate | PW | 0 | 0 | 4 | 2 | 0 | | 5 |
| | EL | 0 | 5 | 4 | 3 | 3 | | |
| **Sample 3** | PW | 0 | 0 | 4 | 2 | 0 | | 5 |
| | EL | 0 | 0 | 5 | 0 | 0 | | |
| Duplicate | PW | 0 | 0 | 4 | 1 | 0 | | 4 |
| | EL | 1 | 0 | 3 | 0 | 0 | | |
| **Sample 4** | PW | 3 | 2 | 1 | 2 | 0 | | 5 |
| | EL | 2 | 5 | 5 | 2 | 3 | | |
| **Sample 5** | PW | 0 | 0 | 0 | 0 | 0 | | 2 |
| | EL | 0 | 2 | 2 | 0 | 0 | | |
| **Sample 6** | PW | 0 | 0 | 2 | 1 | 0 | | 3 |
| | EL | 0 | 3 | 0 | 1 | 0 | | |
| **Sample 7** | PW | 1 | 0 | 0 | 0 | 0 | | 2 |
| | EL | 1 | 0 | 2 | 0 | 0 | | |

*Figure 9: Results of AGI presented as pT-values for pore waters (PW) and eluates (EL) for each laboratory and sediment. Sediments 1-3 were introduced as hidden duplicate samples.*

*Table 6: Minimum and maximum inhibition rates % in D1 obtained by AGI in different laboratories 1-6 for the sediments**

| Sediment | | Inhibition rate % Min | Inhibition rate % Max |
|---|---|---|---|
| **Sample 1** | PW | -47 | 47 |
| **Sample 1** | EL | -43 | 80 |
| **Sample 2** | PW | -89 | 30 |
| **Sample 2** | EL | -67 | 89 |
| **Sample 3** | PW | -93 | 32 |
| **Sample 3** | EL | -108 | 93 |
| **Sample 4** | PW | -23 | 91 |
| **Sample 4** | EL | 26 | 93 |
| **Sample 5** | PW | -418 | -31 |
| **Sample 5** | EL | -142 | 49 |
| **Sample 6** | PW | -118 | 52 |
| **Sample 6** | EL | -36 | 22 |
| **Sample 7** | PW | -228 | 58 |
| **Sample 7** | EL | -44 | 50 |

*laboratory 3 did not join the AGI round robin.

When applying the valid criteria for sediment assessment (BfG, 2011), sediment 1, 2, 3 and 4 should be treated as material that is hazardously contaminated. Sample 6 is classified as critically contaminated and sediments 5 und 7 are classified as material that is only little contaminated (Figure 9).

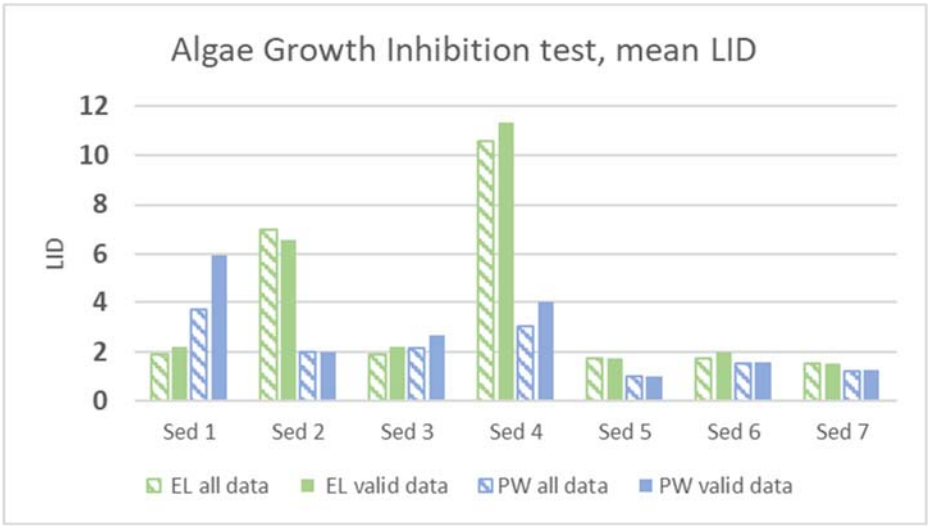



*Figure 10: Mean of LID obtained in the Algae Growth Inhibition test; EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data. (taken from Ratte, 2020).*

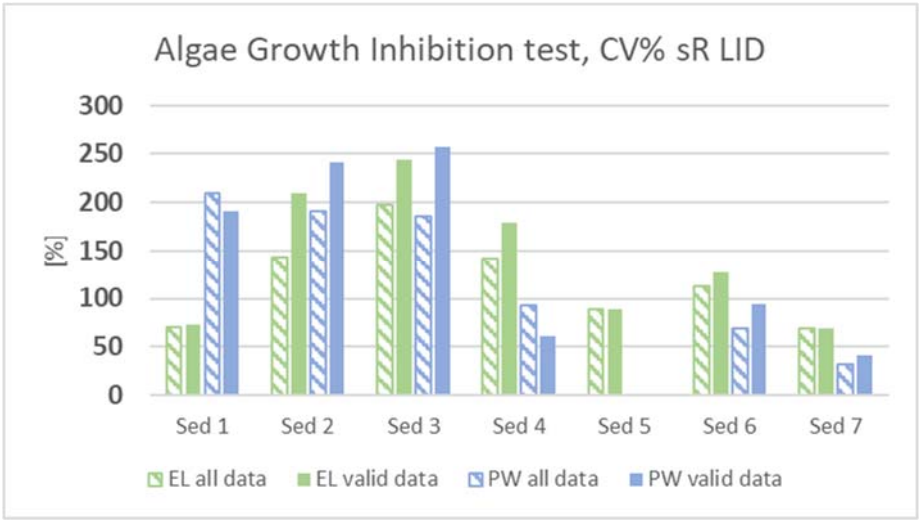



*Figure 11: CV% for reproducibility of LID obtained in the Algae Growth Inhibition test; EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data. (taken from Ratte, 2020).*

Mean values for the LID are visible in Figure 10. Highest mean values were obtained for sediment 2 and 4 for the eluates and sample 1 and 4 for the pore waters. Lowest mean values were obtained for sediment 7 for the eluates and for sediment 5 for the pore waters (Figure 10). The coefficient of variation revealed a variation around or above 200% for samples 1, 2, and 3. For samples 4 and 6 the CV% still exceeds 100% and for sample 5 and 7 CV% are around 50% (Figure 11).

Mean inhibition values above 20% were obtained for sample 1, 2, 3, 4 and 7 for the eluates and for sediment 1 and 4 for the pore waters (Figure 12). Except for sample 4 in the eluates and sample 4 and 5 in the pore waters the normalised width (NW) shows values above or equal 2 meaning “that the true mean will be within a range of twice the calculated mean” (Ratte, 2020).

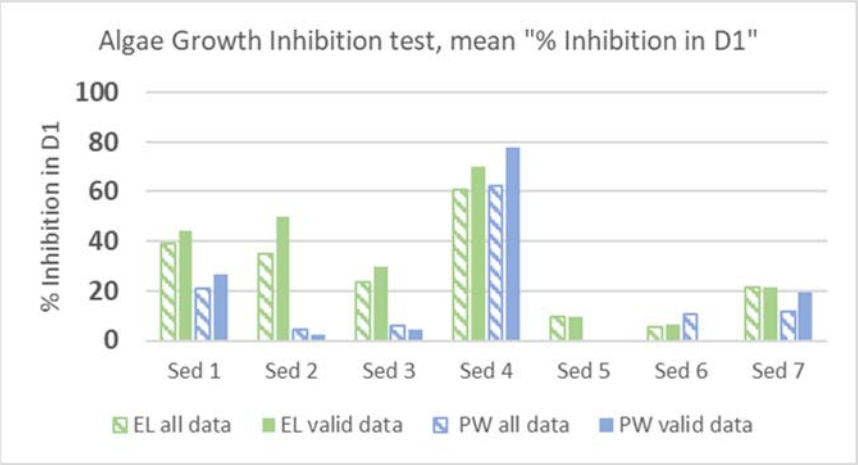



*Figure 12: Mean of inhibition rate (%) in D1 obtained in the Algae Growth Inhibition test, EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data.*

The intra-laboratory variability for the AGI was assessed by analysing duplicates for sediments 1-3. In four samples out of 30 duplicate samples LIDs were found that differ at least two dilution steps, in further six samples LIDs differ not more than one dilution step and identical LIDs were found in 20 samples. Due to statistical reasons no CV could be obtained for the LID data, but on the inhibition rates. CV% for the D1 show that CV% was around 20% for EL samples and between 0% and 280% for PW samples (Ratte, 2020).

| Sample | | LBT Lab 1 | LBT Lab 2 | LBT Lab 3 | LBT Lab 4 | LBT Lab 5 | LBT Lab 6 | FWAT Lab 1 | FWAT Lab 2 | FWAT Lab 4 | FWAT Lab 5 | FWAT Lab 6 | Biotest battery final result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pT | pT | pT | pT | pT | pT | pT | pT | pT | pT | pT | Toxicity Class |
| **Sample 1** | PW | 0 | 2 | 3 | 0 | 1 | 1 | 1 | 2 | 4 | 1 | 0 | V |
| | EL | 2 | 2 | 4 | 0 | 4 | 0 | 2 | 1 | 0 | 1 | 0 | |
| Duplicate | PW | 0 | 2 | 3 | 0 | 1 | 1 | 1 | 5 | 4 | 1 | 0 | |
| | EL | 2 | 2 | 4 | 0 | 3 | 1 | 1 | 2 | 0 | 2 | 0 | |
| **Sample 2** | PW | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 4 | 0 | 0 | V |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2 | 4 | 3 | 3 | |
| Duplicate | PW | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 4 | 2 | 0 | |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 5 | 4 | 3 | 3 | |
| **Sample 3** | PW | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 4 | 2 | 0 | V |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 5 | 0 | 0 | |
| Duplicate | PW | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 4 | 1 | 0 | |
| | EL | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 3 | 0 | 0 | |
| **Sample 4** | PW | 2 | 1 | 4 | 0 | 2 | 0 | 3 | 2 | 1 | 2 | 0 | V |
| | EL | 2 | 1 | 3 | 0 | 1 | 1 | 2 | 5 | 5 | 2 | 3 | |
| **Sample 5** | PW | 2 | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | II |
| | EL | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 2 | 2 | 0 | 0 | |
| **Sample 6** | PW | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | III |
| | EL | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 0 | 1 | 0 | |
| **Sample 7** | PW | 0 | 1 | 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | III |
| | EL | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | |

*Figure 13 Results of biotest battery based on officially assessment scheme (BfG, 2011). Last column reflects final result of all tests.*

## Results obtained by the whole biotest battery

As stated above, assessment of sediments by means of biotest data should follow requirements provided by BfG (2011). Considering the scheme in Table 1, the samples will be classified due the highest pT-value of single tests.

Therefore, sediment 1,2,3 and 4 are classified as sediments that are hazardously contaminated. Sample 6 and 7 are classified as critically contaminated and sediment 5 is classified as material that is only little contaminated (Figure 13).

## Correlation of test results to site conditions

There have been significant relationships identified between the endpoints of ecotoxic tests (percentages in D1 and pT-values) and the oxygen content before aeration, conductivity, ammonia and total nitrogen content. Most significant relationships were obtained for the LBT test results from the eluates. The oxygen content before aeration is negatively correlated to toxicity while conductivity, ammonia and total nitrogen content is positively correlated. Similar results are seen for the AGI in results from pore waters for the inhibition rates in D1. Generally, correlation coefficients are low (Table 7) and correlation of toxicity and nutrients should be seen tentatively since methods of nutrient analyses differ between the laboratories.

*Table 7: Pearson's correlation coefficient regarding ecological test results and side parameter analysed in pore waters and eluates. Only significant (alpha<0.05) relationships are shown*

| **Parameter** | **LBT** | | | | **AGI** | | | |
|---|---|---|---|---|---|---|---|---|
| | **PW** | | **EL** | | **PW** | | **EL** | |
| | *% D1* | *pT-value* | *% D1* | *pT-value* | *% D1* | *pT-value* | *% D1* | *pT-value* |
| **Oxygen before aeration** | - | - | -0.43 | -0.41 | -0.3 | - | - | - |
| **Conductivity** | -0.37 | - | 0.51 | 0.39 | 0.39 | - | 0.46 | 0.33 |
| **Ammonia** | - | 0.3 | 0.42 | 0.5 | 0.4 | 0.38 | 0.40 | - |
| **total Nitrogen** | - | - | 0.5 | 0.41 | 0.44 | - | - | - |

# Discussion

Although much effort was made to synchronise sampling and sample handling for the round robin test, and intra-laboratories reproducibility for the LBT was satisfying, test results show high variability between results from different laboratories. This is visible for the LBT and for the AGI as well as for the physico-chemical parameters of eluates and pore waters. For the AGI single samples show a broad range of toxicity, that range from no toxicity at all to high toxicity in sample 1, 2, 3, and 4 and from no toxicity to moderate toxicity in sample 6 depending on the laboratory (Figure 9). The broad range of responds is also documented by inhibition rates within dilution stage 1 (D1), where inhibition but also stimulation is seen for the same samples (Table 5). Similar effects are visible in test results obtained by the LBT, although toxicity range for single samples is not as broad as found in the results of the AGI. Nevertheless, results vary between no toxicity at all and elevated toxicity in sample 1 and 4 and no toxicity and moderate toxicity in sample 7.

For both tests there is no systematic difference between the results obtained from the pore waters or from the eluates. The reproducibility was assessed for the LID (that could be translated to pT-values) and the inhibition rates in D1. For the AGI the reproducibility of the LID was poor and exceeded 100% in most samples while NW were up to 10 (Figure 7 and Annex 3). For the LBT the reproducibility was slightly better and only some CV% exceeded 100% and NW were up to 10 (Figure 7, and Annex 3).

The laboratories followed the same protocols for preparation of test material (pore waters and eluates) and applied the same methods for running the AGI and LBT. Whether smaller methodological differences like different sources of

DIN EN ISO 11348-2 conform liquidly dried bacteria in laboratory 4, different measurement devices and the miniaturised LBT of laboratory 3 represent main reasons for different results for the same sediments needs to be further discussed. One assumption is that test organisms might be more or less sensitive in different labs. The sensitivity of test organisms might be detected by their response to the reference substance. While laboratory 3 who applied as only participant a miniaturised LBT was found to overestimate toxicity deduced from the LBT compared to the other laboratories (Annex 3), the response of the bacteria to the reference substance 3,5-Dichlorphenol 5 mg/L was moderate compared to the other laboratories. Although differences of test material properties (Table 4) need to be handled with care since laboratories applied different methods and used different devices, those differences from same sediments might indicate that variability of biotest results could be related to differences of test material. Correlation analyses with test material properties point to dependence of test results on conductivity, ammonia concentration, pH-value and dissolved organic carbon for the AGI and conductivity for the LBT (Martens, 2017). When running correlation analyses for the present data of the round robin test (Table 3), similar relationships are visible and significant relationships are found for the LBT between the inhibition rate in D1 in the eluates or pore waters and ammonia, conductivity, total nitrogen, and the oxygen content before aeration (Table 7). For the AGI also a significant relationship between inhibition rates in D1 and ammonia, conductivity, total nitrogen, and the oxygen content before aeration was found. The relation with the oxygen content before aeration is negatively related to the endpoints. This implies that redox conditions prior to elutriate preparation may have had an effect on different biotest responses. Toxicities in LBT have, for example, been attributed to sulfide species.

It is assumed that bioassay responses reflect the integrated impact of the test organisms' exposure towards inhibiting substances. Identification of cause-effect relationships for sediment toxicity in most cases requires substantial effort, as different substances mostly contribute to varying degrees to this toxicity. Not only new and historic contaminants but also naturally inhibiting chemicals such as toxins, or confounding factors may cause inhibition. Especially the presence of toxins but also nutrient concentrations, organic carbon content, and further confounding factors vary seasonally and may be responsible for short-termed effects. However, in this study, all laboratories were supplied with homogenised sediments from the same location and sampled at the same time. Test materials – if prepared in an identical way – would be expected to show comparable toxicities. In contrast to this, testing resulted in a broad range of ecotoxicological responses, and these, following the current decision framework, resulted in classifying two very different sediments as moderately toxic: Sediment 7 which was considered to be relatively strongly contaminated, and the reference sediment 6 that had been derived from Holocene layers and was without any detectable chemical contamination (see Table 3). This puts the present biotest assessment scheme into question.

Within the statistical assessment of the round robin test done by Ratte (2020) it is discussed whether the AGI and the LBT are suitable to assess toxicity of different natural sediments. It is stated that regarding the LIDs as endpoint both AGI and LBT showed highly variable, but nevertheless consistent results regarding the general toxicity level measured, except for sediment 2. It is also stated that the precision of the obtained mean LIDs is very low and thus the reliability of a single result from a single laboratory also is found to be very poor, while the results from LBT were less variable than those from AGI. Ratte (2020) concluded that this might be due to the overall lower sensitivity of the LBT and / or to lower degree of interferences due to confounding factors.

In most cases only one laboratory provides biotest results for the assessment of sediments for sediment management decisions. General tendencies by means of (geometric) means of single test results from different laboratories, like

applied for the statistical evaluation of the round robin test, could therefore not be used. At present, the valuation scheme in Germany (Table 1) does not take temporal variability or the low precision of biotest results for sediments into account and only maximal deduced toxicity of the a single biotest decides how to handle the material. This is documented in Figure 13 where all information from the single ecotoxic tests is summarized for the final result. It is still not clear which role the contamination level of sediments and the bioavailability of the contaminants play in the response of ecotoxic test systems in relation to confounding factors. A recent study with sample sizes >400 point to only very low correlation (with correlation coefficients between 0.5 and 0.6) of AGI test results (D1 and pT-value) with the contaminants analyzed and slightly higher (r: 0.55-0.8) for confounding factors like ammonia, dissolved organic carbon, pH and conductivity (Martens, 2017). For the LBT no significant correlation with any contaminant was detected (ibid). Also studies on pesticide contaminated sediments found ecotoxicological response of biotest organisms that could not be explained by the detected contamination levels (Casado-Martinez et al. 2020). There are probably unknown interactions of compounds present in the environment or compounds that have not been detected yet in sediments and ecotoxicological testing could help assessing possible hazardousness. However, responses of ecotoxicological test systems could also – at least partially – derive from other factors than hazardous contaminant concentrations (see above).

While it is still not clear on which factors organisms used in ecotoxic testing react and test results seem to only provide low precision, an integrated assessment scheme for the assessment of sediments is highly desired. Promising prospects for an integrated assessment of biotest data could be the evaluation on battery level where all single test results will be considered. Examples for such approaches could be found in European countries like Italy (Regoli et al., 2019) where the assessment of biotests for sediment assessment is established within the frameworks for longer times than in Germany (Heise et al., 2019).

## Conclusions

For the here presented round robin test, the reproducibility of test results was poor, and the variability of results obtained from different laboratories was up to 200% for the AGI and up to 100% for the LBT.

The standardised sampling and homogenisation routines did not lead to a satisfying reproducibility of AGI and LBT results between different laboratories but might improve the repeatability of test results within the same laboratories as seen for the hidden duplicate samples from the LBT. However, further studies to improve understanding of biotest responses and reliability of biotest results are required.

Up to now, existing frameworks for sediment assessment do not consider temporal variability. This should be adjusted in future regulations.

According to our study, biotest results, run under the current instructions, could not be used to assess the toxic potential of sediments in a reliable way and the question posed in the title of the here presented study remains yet unsolved. However, tendencies for differences in toxicity could be seen by integration of different test results. This should be taken into account when using data for sediment management decision, following European countries like Italy for the development of new assessment schemes and frameworks.

# Chapter 3 Interlaboratory comparison of bioassay results: Investigating the impact of methodological differences

By Sonja Faetsch (HAW), Susanne Heise (HAW), Nicole Schlichting (IDN)

## Introduction

The results of the previous two chapters showed, that interlaboratory differences of biotest results from the AGI and LBT can be high. In the Sullied Sediments Project, all sediment samples were well homogenized before they were sent off to the labs. Nevertheless, resulting data on toxicities differed between IDN and HAW. Variations between the labs' procedures regarding elutriate preparations, time between sampling and test start, elutriate processing, but also different sensitivity of algae species were identified as potential causes (chapter 1). All labs of the round robin test (chapter 2) were supposed to follow the given SOPs. These SOPs, however, still leave room for modifications or allow some flexibility in the test procedure. This may have resulted in the differences in especially oxygen concentrations and conductivity of pore waters and elutriates that were measured by the participating labs prior to performing the biotests for the round robin test (see Table 4, Chapter 2).

Because of the differing results of biotests in the Sullied Sediments project (chapter 1), the two labs at HAW and IDN decided to start an additional testing round, looking more closely at the inhouse-methods and the differences between them, with the aim to identify causes for the observed deviations. Elutriates that had been prepared by the respective inhouse method were exchanged between labs. Then each lab carried out its usual bioassay protocol on both, IDN and HAW, elutriates. The aim was to examine whether e.g. elutriate preparation, modifications of the biotesting, measuring devices or the test organisms would substantially influence the outcome of the test.

## Methods

### Sampling

Three sediment samples were taken in June 2019 from water bodies around Hamburg, Germany. Sampling and analysis was repeated in April 2020. Site I was located at Harburg Harbour, site II at Schweenssand and site III at the Dove Elbe, whereby the sampling sites in 2019 and 2020 were not identical due to logistical reasons (e.g. limited access to sites in 2020 due to Corona restrictions). The samples were kept at 4°C at the two laboratories, exchanged by car transport the following day. The difference between the analysis in 2019 and 2020 was such, that the self-imposed requirement was to start at the **same day** in 2019, and at exactly the **same time** in the two labs in 2020.

### Elutriate preparation and biotesting

Both laboratories prepared elutriates according to their protocols, as described in chapter one, and exchanged them. Transport took place in cooled boxes and the biotests started on the day of exchange. The laboratories analysed both kinds of elutriates, the self-prepared ones and the ones from the other lab. Elutriates were analysed with the two biotest systems, the AGI (algae growth inhibition test) with *D. subspicatus* (IDN) and *R. subcapitata* (HAW) and the LBT (luminescence bacteria test). Descriptions of test procedures are described in chapter one.

# Results and Discussion

## Comparability of Elutriate preparation

While HAW prepares elutriates by resuspending 5 $cm^3$ of wet sediment in 20 ml of water, IDN follows the protocol of Floeter (2009) in which elutriates are prepared from a fixed particle/water ratio: fresh sediment is suspended in such an amount of water, that one part of sediment (dry weight) is complemented with 3 parts water (including the porewater), unless the water content exceeds 75% in which case no additional water is added and the sediment is elutriated in itself. The different procedures lead to fundamentally different water content of the resulting elutriates with IDN-elutriates always having a higher sediment concentration, as also indicated by the higher conductivity and lower oxygen concentration (Table 8).

Slight or no turbidity was observed in HAW-elutriates after centrifugation and before the test and no further processing such as centrifugation was considered to be necessary. The IDN-elutriates showed, depending on the sample, stronger turbidity in the centrifuged supernatant which seemed to increase with time. Therewith the time span between elutriate preparation and analysis could be critical. The first comparison-series in 2019 (Figure 6) did not pay attention to this factor. In the second comparison-series in 2020, care was taken to synchronize the times when elutriate preparation and bioassays were carried out in the two labs.

All tests were carried out with a 2-fold dilution series whereby HAW always measured the original elutriate and 4 dilutions, and IDN diluted until the inhibition dropped below 20 %. In the following graphs, inhibition data produced by HAW and IDN always comprise the results from both elutriates that were tested per lab. The data were made comparable by calculating the percentage of pore water in each tested dilution step on the basis of sediments' water content and the respective protocol for how the elutriates were set up.

*Table 8: comparison of dry weights of sediments I to III, and of the properties of samples, eluted according to the volumetric (HAW) and to the gravimetric (IDN) method, measured by Lab 2*

| | parameter | Volumetric method | Gravimetric method |
|---|---|---|---|
| sediment I | Dry weight in % | 15 | |
| | Conductivity [µS/cm] | 395 | 526 |
| | Oxygen concentration [mg/l] | 9,67 | 2,06 |
| | Turbidity of elutriate | slight | strong |
| | Range of pore water fractions in the dilution series (exposed concentration) | 0.5 – 14.0 % | 3.3 – 80 % |
| sediment II | Dry weight in % | 20,8 | |
| | Conductivity [µS/cm] | 833 | 1568 |
| | Oxygen concentration [mg/l] | 9,78 | 2,14 |
| | turbidity | slight | slight |
| | Range of pore water fractions in the dilution series (exposed concentration) | 0.5 - 13.2 % | 3.3 - 80 % |
| sediment III | Dry weight in % | 70,1 | |
| | Conductivity [µS/cm] | 238 | 751 |
| | Oxygen concentration [mg/l] | 10,01 | 7,87 |
| | turbidity | no | no |
| | Range of pore water fractions in the dilution series (exposed concentration) | 0.2 - 5.6 % | 0.4 – 11.4 % |

## Algae Growth Inhibition Test

Figure 14 depicts the results of AGI tests performed at IDN and HAW. The differently prepared elutriates are indicated with squares (HAW-method) and circles (IDN-method). Results from the two laboratories are indicated in green (HAW, biotest organism *R. subcapitata*) and purple (IDN, biotest organisms *D. subspicatus*). A third order polynomial interpolation of the HAW and IDN data set was carried out to check whether the data from differently prepared elutriates would fit together. Goodness of Fit ($R^2$) for the interpolation were 0.7093 (HAW-Testing) and 0.8106 (IDN Testing) for sample I, and 0.8668 (HAW-Testing) and 0.8289 (IDN-Testing) for sample II. These results indicate that

- ➔ for samples I and II, the different elution procedures do not result in fundamentally different inhibition data of the respective test organisms. With reference to the pore water percentage in the tested volume, combined data from the two elutriates form a comprehensible dose-response relationship.
- ➔ Results from the two laboratories differed, and most distinctively with sample I. With increasing pore water fraction in the tested sample, stimulation increased at IDN, while inhibition rose at HAW. As with sample I, results for sample II from Lab 1 (HAW) indicated a toxicity at lower pore water concentrations than results from Lab 2 (IDN).

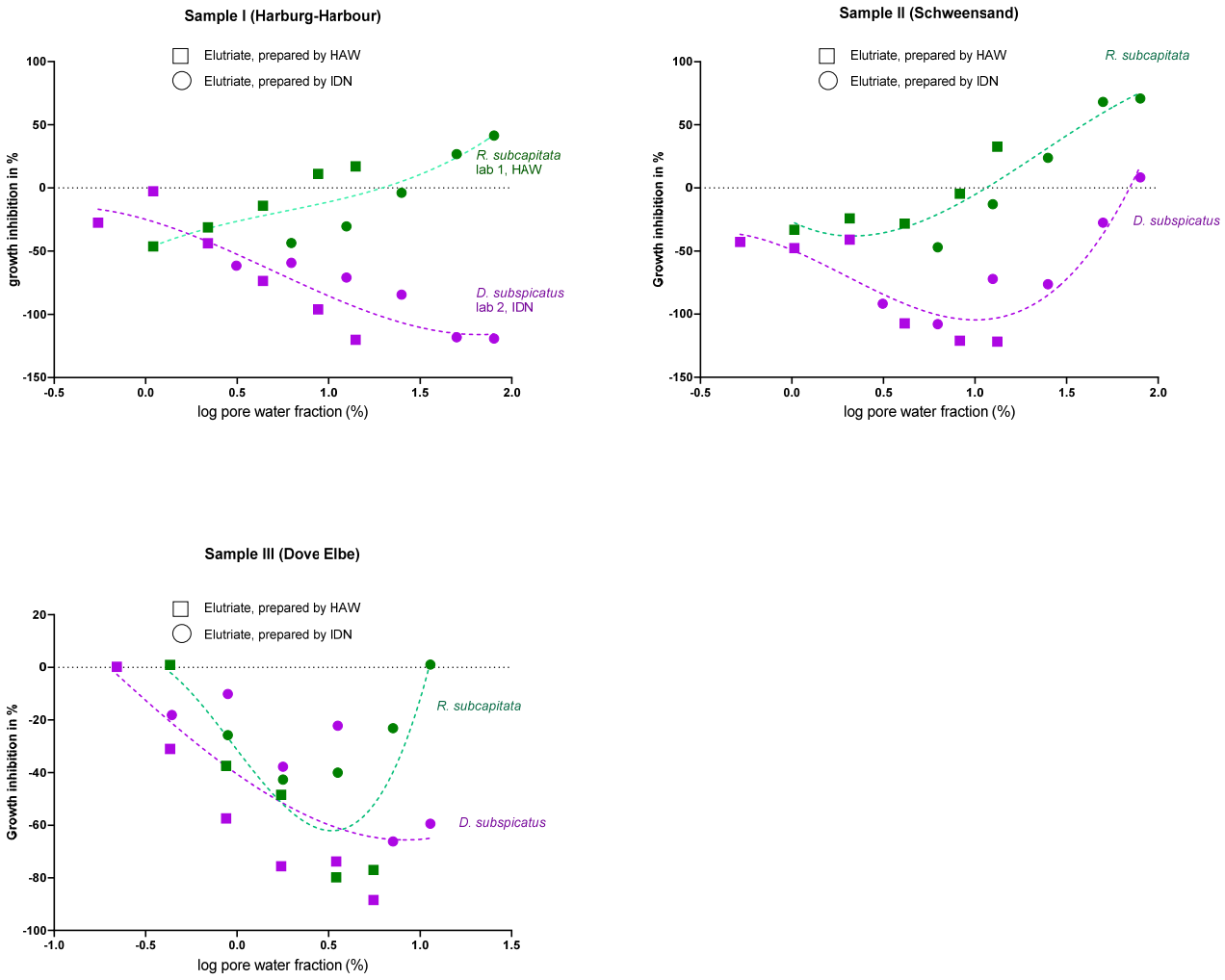



*Figure 14: Results of the algae growth inhibition test (AGI) of the comparison-series in 2019. Laboratory IDN applied D. subspicatus, HAW applied R. subcapitata. Both labs measured elutriates prepared by the HAW method (squares) and the IDN method (circle). Mean values of algae growth inhibition determined by fluorescence measurements of elutriates compared to unexposed control cultures. From top left to bottom: Site 1: Harburg Harbour, Site 2: Schweenssand, Site 3: Dove Elbe. Green: R. subcapitata, (Lab 1, HAW). Purple: D. subspicatus (Lab 2, IDN).*

With sample III, differences between labs were not as obvious. Due to the high dry weight of this sediment, the different eluting procedures resulted in a similar and low range of pore water fractions in the tested dilutions (Table 8). Inhibition in all samples was zero or negative and no clear trend could be observed.

➔ With sample 3, differences between labs in the AGI test become more pronounced with higher fraction of pore water in the tested material. However, the high dry weight suggests, that this sample was sandy and thus probably little contaminated.

Differences between the labs' results, apparent with sample I and II, seem to be due to the test procedures rather than the elutriate preparation. Both labs carry out the biotest according to a DIN Norm, but from different years (DIN EN ISO 8692:2012 at Lab 1 and DIN 38 412-L33 (1991) at Lab 2). Required media for the tests only differ in 2 aspects between the norms: In the final concentration of $NaHCO_3$ (50 mg/L in DIN EN ISO 8692; 300 mg/L in DIN 38412-33) and of $FeCL_3$*6 $H_2O$ (64 µg/L in DIN EN ISO 8692; 80 µg/L in DIN 38412-33) in the test. In addition, Lab 1 performs the test in 24-well microtiterplates, while Lab 2 uses Erlenmeyerflasks. Even though the influence of subtle other differences in the test procedure can currently not be excluded and should be subject to another interlaboratory study, the major distinction between the two labs is the applied test organism: While IDN performs the test with *D. subspicatus*, HAW uses *R. subcapitata*.

According to the guidelines DIN 8692:2012 and OECD 201, both unicellular green algae species *R. subcapitata* and *D. subspicatus* are eligible for use in algae growth inhibition assays. The round robin test published in DIN EN ISO 8692:2012-06 shows only a slight difference between these two species, with *R. subcapitata* having a lower $EC_{50}$ for 3,5-Dichlorphenol (DCP). The higher sensitivity towards 3,5-DCP was confirmed in this interlaboratory comparison (Figure 15). The fact, that in the round robin test, reaction towards potassium dichromate was comparable between the two species, shows, that sensitivities towards different substances may be higher, smaller or similar between different test species.

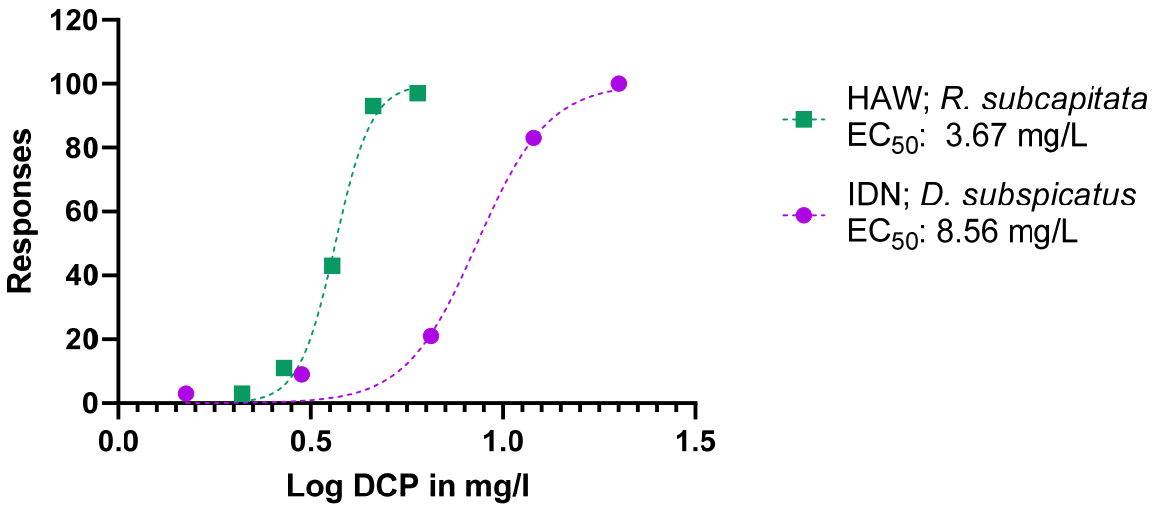



*Figure 15: Dose response curve of R. subcapitata and D. subspicatus and $EC_{50}$ values for 3,5-DCP. Data were compiled at the respective labs.*

In the literature, most comparisons of algae species sensitivities have been done with single substance toxicity studies. Exposing *R. subcapitata* and *D. subspicatus* to various substances showed that sensitivity varied depending on tested chemicals. It could not be stated that one species is generally more or less sensitive than the other one. E.g. *R. subcapitata* was more sensitive than *D. subspicatus* to 3 metal compounds (chromium, copper and zinc) and 3 herbicides (oxyfluorphene, pendimethaline, atrazine) (Rojíčková and Maršálek 1999) and the other way round in case of acrylamide (Zovko et al., 2015). Becker et al. (2007) detected a 4 times higher sensitivity of *R. subcapitata* in

comparison to *D. subspicatus* towards heavy metal contaminated municipal waste incineration ash, and a 5 times higher sensitivity towards waste wood, also with a high copper content.

When comparing the results from the two labs for all 6 sampling campaigns (Figure 16), the following can be observed:

- Results for the AGI test differ between the two labs for the first 3 campaigns, in that Lab 1 (HAW, *R. subcapitata*) measured high toxicities (more than 80 % at more than 8 % of pore water fraction, at site 1), while bioassays of Lab 2 (IDN, *D. subspicatus*) observed low toxicity at the same pore water concentrations and increased responses only at higher pore water concentrations at Site 1 (campaign 1 and 3).
- Toxicities measured for sites 2 and 3 are similar and low for both laboratories for the campaigns 2, (3) 4, 5 and 6.

These data need to be seen in the light of a persisting contaminant gradient in the Elbe Estuary. Concentrations of historical pollutants decrease along the estuary towards the mouth of the river due to upstream sources and transport of contaminated sediment (Wetzel et al., 2013; Heise et al. 2007a,b). Site 1, Stover Strand, is the most upstream sampling site and the most exposed to the upstream emissions. At this site, the algae test with *R. subcapitata* shows high toxicities from October 2017 to Summer 2018. Site 2 is located in the Hamburg Harbour area where contaminant concentrations are lower due to tidal mixing processes. Toxicities here are high in October 2017 and Summer 2018, Site 3 (Wedel) is further downstream, least exposed to historical contaminants, and – apart from October 2017 – in line with the results obtained with *D. subspicatus*. In the light of the cited comparisons between the two species, *R. subcapitata* seems to reflect the contaminant gradient over the Elbe estuary – at least regarding historical substances with upstream origin. Further studies should be carried out to exclude any confounding influences on the toxicity which may have led to this outcome. If the sensitivities of this algae species indeed reflects the Elbe-typical contamination pattern, a discussion should start on the potentially better suitability of this test organism for the identification of hazardous sediments in the Elbe region.

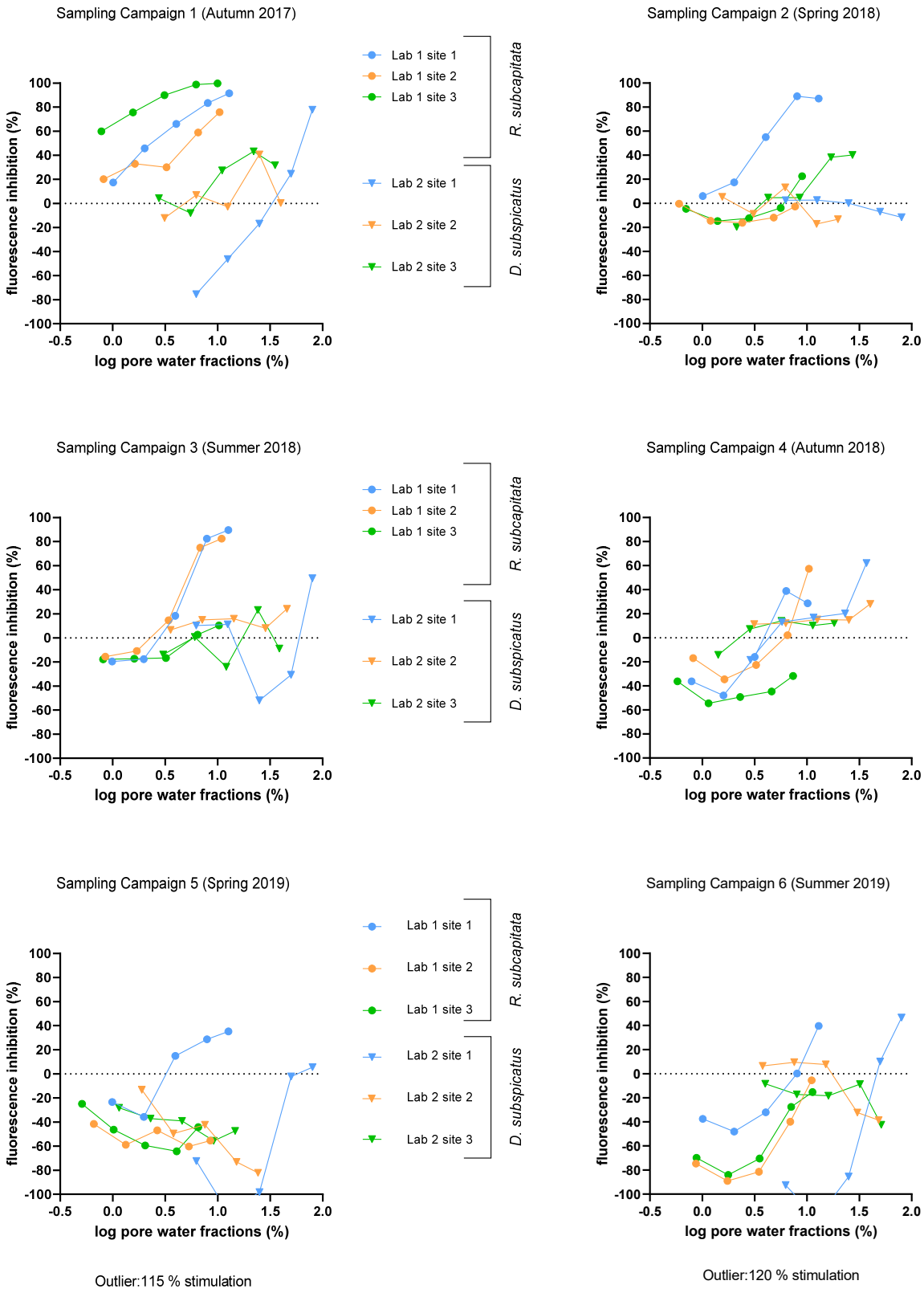



*Figure 16: Compiled AGI-Inhibition data for German Sullied Sediment samples (all campaigns; Lab 1: HAW; Lab 2: IDN)*

## Luminescence Bacteria Test

Figure 17 shows the results of the first interlaboratory comparison in June 2019 when the samples from the 3 sites were tested with luminescence bacteria. Inhibition values are plotted as a function of pore water concentration in the test batch (Pore water fraction in %). For easier comparison, a non-linear regression line was added to the data (polynomial, 3$^{rd}$ order) not implying statistical significance.

Comparison of elutriates that were prepared in the 2 labs: In most samples, inhibition data of elutriate testing could be fitted to one regression line that connected all data measured by one laboratory, independent on where the elutriates had been prepared. The only exception was Sample 1 where elutriates from HAW and IDN differed when measured by IDN. These data imply that the influence of elutriate preparation on the outcome was low for those samples that were measured here.

Comparison of results between labs: Here, the bioassays gave different results. For sample I, results were consistent, ranging from 20 to 91 % inhibition at HAW, and 13 to 97 % at IDN. The samples II and III resulted in significantly higher toxicities when measured at HAW than when measured at IDN. Especially the sample III from Dove Elbe would have been assessed to be highly toxic in the HAW-test and not toxic at IDN. As elutriate preparation does not seem to lead to different results (see above), the reasons for this deviation must lie in the procedures of respective laboratory. As one reason, the time delay in the start of the biotesting was identified. While IDN ran the LBT in the morning, HAW started in the evening when the cooled sample was tested without further centrifugation.

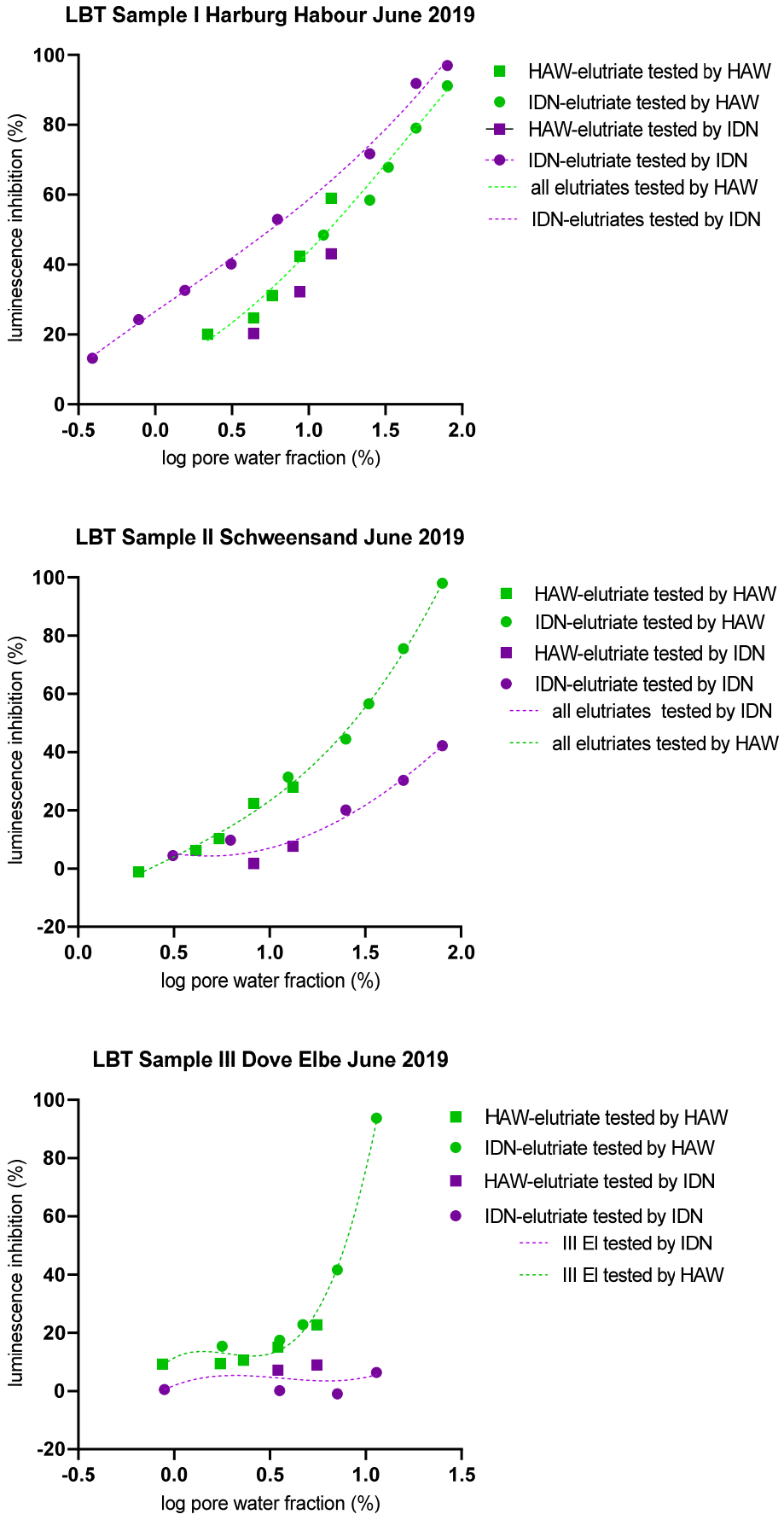



*Figure 17: Results of luminescence bacteria tests of first comparison-series between two labs (Lab 1 HAW and Lab 2 IDN)) in June 2019. Mean values of bacterial luminescence inhibition in elutriates compared to unexposed control cultures. From left to right: Site 1: Harburg Harbour, Site 2: Schweenssand, Site 3: Dove Elbe. Green: Tests conducted at Lab 1 HAW, squares: elutriate prepared by HAW, circles: elutriate prepared by IDN. Purple: Tests conducted at Lab 2, IDN, squares: elutriate prepared by HAW, circles: elutriate prepared by IDN. Dark purple: Results without colour correction, light purple: colour corrected values.*

The test was repeated by Lab 1 (HAW) with frozen elutriates from samples II and III in December 2019. The samples were thawed and measured in replicates a) after thorough mixing, b) after centrifugation for 5 minutes at 10 000 g.

Results are depicted in Figure 18 in comparison with the former results from June 2019 directly after sampling. In both tests, inhibition was significantly reduced compared to the first measurements in June. For Sample II which derives from Schweenssand and was a fine grained sample, centrifugation lowered the toxicity even more compared to the shaken sample (by 10 to 20 %), while there was no significant difference (alpha: 0.05) between the toxic responses of the shaken and the centrifuged sample for the sandy material of Site III. However, toxicities were still higher than those measured by IDN in June. At this time, no conclusion could be drawn from these data, other than treatment (shaking versus centrifugation) can have an influence on the measured toxicity, probably due to colloids or particles which may have been formed.

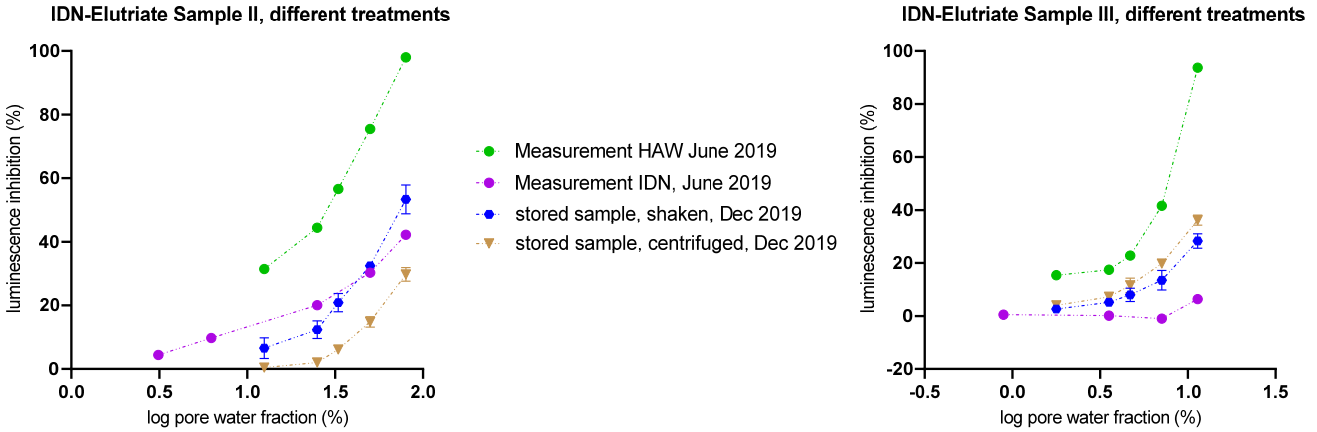



*Figure 18: Results from repeated LBT on frozen elutriates after vigorous shaking or centrifugation by HAW, compared to original data from June 2019 of both labs*

Whether the high toxicity of HAW or the lower values of IDN reflected the “correct” toxicity of the samples was not possible to answer. Assuming, that particle/colloid formation with time may have had an influence on the samples and the bioassay outcomes, the second comparison-series in April 2020 focussed on identical processing times.

The results for this comparison are presented in Figure 19. Results were similar between the two labs with toxicities measured by HAW being slightly higher by 5-18 percentage point, with the exception of Sample II, in IDN elutriate, when IDN toxicities were slightly higher. In opposite to the interlaboratory exercise in June 2019, differences between elutriate preparations were more pronounced, especially with sample III: Here, elutriates prepared by the gravimetric method (IDN) showed lower toxicities compared to the volumetric method (HAW). Toxicities measured by the two laboratories on the respective elutriate were almost identical.

In this interlaboratory comparison carried out in April 2020, variation between the two labs was below 20 percentage points. This is in the range of the deviation which the producer of the luminescence bacteria predicts for different batches of bacteria with regard to the positive control of potassium dichromate.

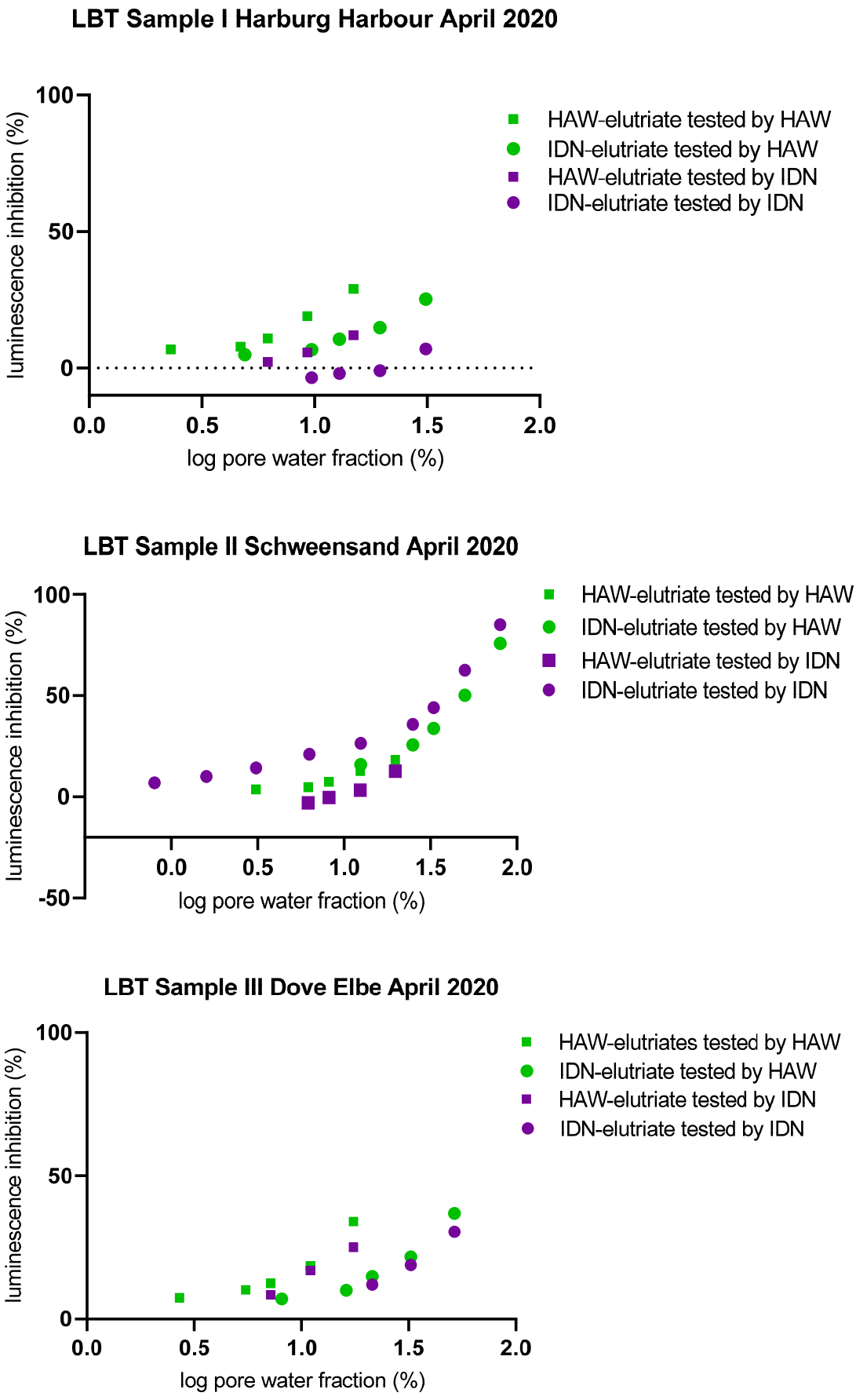



*Figure 19: Results of luminescence bacteria tests of second comparison-series between two labs (Lab 1 HAW and Lab 2 IDN) in April 2020. Mean values of bacterial luminescence inhibition in elutriates compared to unexposed control cultures. Sediment density considered for HAW samples. From top to bottom: Site 1: Harburg Harbour, Site 2: Schweenssand, Site 3: Dove Elbe. Green: Tests conducted at Lab 1 HAW, squares: elutriate prepared by HAW, circles: elutriate prepared by IDN. Purple: Tests conducted at Lab 2, IDN, squares: elutriate prepared by HAW, circles: elutriate prepared by IDN.*

# Conclusions

Preparation of elutriates needs stronger focus. While in June 2019, the different procedures applied by the two labs, seemed to have little effect, differences were more pronounced in the second round of experiments (April 2020). The impact of sediment and elutriate handling prior to the test thus seem to be different with different sediment samples. The fact that partitioning of chemical substances between the sediment matrix, the pore water and the atmosphere depends on a number of properties such as temperature, redox potential, mechanical stirring, water/sediment ratios etc. is well known and has been shown by e.g. Ebert et al. (2012). They demonstrated for soil that the measured toluene content changed with handling of samples like repeated stirring and thus with redox conditions. This was also observed with sediments (unpublished data). As elutriates need to be oxic for the bioassays, changes in redox conditions during elutriation (and potentially aeration) will result in chemical reactions, such as formation of complexes and colloids, and varying desorption of contaminants. Turbidity that may arise from transport, storage or aeration of elutriates could demand further centrifugation steps, that may affect the exposed contaminant concentration in the tests.

Consequently, the critical steps of the elutriate preparation need to be harmonized. These may differ and depend on sediment properties and the kind and age of contamination. Further and probably extensive studies to give robust answers with regard to harmonization requirements seem to be necessary.

An additional factor seems to be the storage time of elutriates. Differences between the two labs, that were still apparent in the 2019 interlaboratory comparison, could be reduced to below 20 % in April 2020, when the time of testing was synchronized. A variation of toxicity by less than 20 %, measured by two different labs, is small in the light that testing procedures are different: While IDN carries out the test in vials, HAW carries it out in 96 well plates. Different devices are used. The individual test protocol may lead to principally higher toxicity signals in the HAW testing. If sensitivities and response ranges are known and consistent, they can be addressed in an assessment scheme that leaves room for laboratory specific adaptations in the interpretation of toxicity data as they are discussed in the Sullied Sediments project.

Differences in the AGI between labs in the interlaboratory tests and when measuring the Sullied Sediment samples were much higher than with the LBT test in the 2$^{nd}$ interlaboratory test. Taking into account literature-documented differences in the sensitivity of the two algae species, the demonstrated deviations in the dose response curve of the positive control, and the reflection of the contamination pattern in the Elbe Estuary by *R. subcapitata*-test results, the different sensitivities of the two algae species could be - at least partly - responsible for the high differences of results between labs. Further studies are needed in order to test this hypothesis.

# Chapter 4: Summary and Conclusion

The authors of this study conclude that:

- In order to improve reliability and comparability of ecotoxicological tests, the most critical methodological steps need to be identified, communicated and strictly controlled during test procedures.
- The interlaboratory LBT-comparison showed that the extent to which sample preparation and handling affects the outcome of the test varies with sediment, and can lead to substantially different outcomes in the bioassays. With synchronized and harmonized procedures, interlaboratory variations were between 5 and 18 percentage points and thus almost matched the in-house variability arising e.g. from different batches of luminescent bacteria.
- As it has been done with chemical analytical methods, interlaboratory comparisons of ecotoxicological testing of sediments should be carried out much more often in order to train staff and establish routines.
- A trend towards slightly higher toxicity was observed in the data from HAW compared to data from IDN. These can derive from different testing protocols (such as miniaturized set up versus set up in vials). Systematic differences in response ranges can be taken into account when interpreting the outcome of the bioassays and should be addressed in the assessment scheme for sediment and dredged material which is in development by the Sullied Sediment project.
- An assessment scheme should be developed on the bases of an integration of biotest results. Individual bioassays should not be assigned such far-reaching decision-making power that they determine management decisions. Instead, more importance should be placed on the outcome of biotest batteries and an integrated assessment of the different biotest responses, following a weight of evidence approach. This may result in a decision framework that should be environmentally safer while not being more costly.

# References


Anderson BPM (1977) A 96-hour sediment bioassay of Duluth and Superior Harbor basins (Minnesota) using *Hexagenia limbata, Asellus communis, Daphnia magna,* and *Pimephales promelas* as test organisms. Bulletin of Environmental Contamination & Toxicology, 18: 159-169

ASTM International (2000) Standard guide for collection, storage, characterization, and manipulation of sediment for toxicological testing. In: ASTM Guide 1391-94, annual book of ASTM standards. American Society for Testing and Materials, West Conshohocken, PA

Becker, R., Donnevert, G., Römbke, J. (2007): Biologische Testverfahren zur ökotoxikologischen Charakterisierung von Abfällen. Forschungsprojekt im Auftrag des Umweltbundesamtes.

BfG (2011) BfG-Merkblatt "Ökotoxikologische Baggergutuntersuchung", Ökotoxikologische Untersuchung von Sedimenten, Eluaten und Porenwässern, Bundesanstalt für Gewässerkunde, Koblenz: Stand: September 2011

Casado-Martinez, C., Schneeweiß, A., Thiemann, C., Dubois, N., Pintado-Herrera, M., lLara-Martin, P.A., Ferrari, B.J.D.: Werner, I. (2020): Ökotoxizität von Bachsedimenten. Wasser und Abwasser. 5:55-67.

Chapman PM (1990) The sediment quality triad approach to determining pollution-induced degradation. Science of The Total Environment 97-98:815-825

den Besten PJ (2007) Risk assessment approaches in European countries. In: Heise S (ed) Sediment Risk Management and Communication. Elsevier, Amsterdam (in press),

den Besten PJ, de Deckere E, Babut MP et al. (2003) Biological effects-based sediment quality in ecological risk assessment for European waters. Journal of Soils and Sediments 3:144

Ebert, J., Poltrock, U., Preuß, T., Fauter, K. (2012): Toluol-Bildung und –Abbau im Laufe der Rübenkampagne. Sugar Industry, 12: 769-773.

Floeter, Carolin (2009a): Basisdokument zur Probenvorbereitung von limnischen Sedimenten für die Durchführung von limnischen Biotestverfahren im Auftrag der Hamburg Port Authority, Stand 2009_03_17.

Floeter, Carolin (2009b): Basisdokument zum Grünalgentest mit *Desmodesmus subspicatus* basierend auf DIN 38 412-L33 (1991) und DIN EN ISO 8692 (2005) für die Bewertung von Sedimenten/ Baggergut im Auftrag der Hamburg Port Authority, Stand 2009_03, Modifiziert im August 2018 durch HPA

Giesy JP, Graney RL, Newsted JL et al. (1988) Comparison of three sediment bioassay methods using detroit river sediments. Environmental Toxicology and Chemistry 7:483-498

GÜBAK (2009): Joint Transitional Arrangements for the Handling of Dredged Material in German Federal Coasal Waterways (GÜBAK-WSV) between federal authorities and federal states (Bremen, Hamburg, Mecklenburg-Western Pomerania, Lower Saxonia, Schleswig-Holstein).

HABAB (2017): Handlungsanweisung für den Umgang mit Baggergut aus Bundeswasserstraßen im Binnenland (HABAB-WSV 2017). Bundesministerium für Verkehr und digitale Infrastruktur, Bundesanstalt für Gewässerkunde, Generaldirektion Wasserstraßen und Schifffahrt. Bonn.

Han, Q., Zheng, Y., Qi, Q., Peng, J., Song, J., Guo, J., & Guo, J. (2020). Involvement of oxidative stress in the sensitivity of two algal species exposed to roxithromycin. Ecotoxicology, 29(5): 625–633. https://doi.org/10.1007/s10646-020-02192-6

Heise S, Calmano W, Ahlf W et al. (2007a) Environmental challenges for the Hamburg stretch of the River Elbe and its catchment with regard to the Water Framework Directive. In: Leal W, Hellsten S, Krahn D, Ulvi T (eds) Baltic River Basin Management Handbook. Strategies for Sustainable River Basin Management. pp 47-61

Heise S, Krüger F, Baborowski M et al. (2007b) Bewertung der Risiken durch Feststoff-gebundene Schadstoffe im Elbeeinzugsgebiet Im Auftrag der Flussgebietsgemeinschaft Elbe und Hamburg Port Authority, erstellt vom Beratungszentrum für integriertes Sedimentmanagement (BIS/TuTech) an der TU Hamburg-Harburg. Hamburg. 349 p.

Heise, S., SedNet and Sullied Sediments Partner (2019): Workshop report on sediment classification and management decisions Hamburg Sept. 20-21 2018.

Heise S, Babut M, Casado C et al. (2020) Ecotoxicological testing of sediments and dredged material: an overlooked opportunity? Journal of Soils and Sediments, 20(12): 4218-4228

Juneau, P., Dewez, D., Matsui, S., Kim, S. G., & Popovic, R. (2001). Evaluation of different algal species sensitivity to mercury and metolachlor by PAM-fluorometry. Chemosphere, 45(4–5): 589–598. https://doi.org/10.1016/S0045-6535(01)00034-0

Lee CR, Peddicord RK (1988) Decision-making framework for management of dredged material disposal. In: Salomons W, Förstner U (eds) Environmental Management of Solid Waste – Dredged Material and Mine Tailings. Springer, Heidelberg, pp 324-371

Ma, J, Wang, P, Chen, J. Sun, Y., Che, J. (2007). Differential response of green algae species *Pseudokirchenrialla subcapitata, Scendesmus quadricauda, Scendesmus obliquus*, *Chlorella vulgaris* and *Chlorella pyrenoidosa* to six pesticides. *Polish Journal of Environmental Studies, 16*(6): 847-851

Martens, N. (2017): Analyse ökotoxikologischer Testergebnisse von Hamburger Elbsedimenten in Beziehung zu Schadstoffgehalten und weiteren sedimentbezogenen Parametern. Bachelor Thesis

Normenausschuss Wasserwesen (NAW) im DIN - Water quality - Fresh water algal growth inhibition test with unicellular green algae (ISO 8692:2012); German version EN ISO 8692:2012

Normenausschuss Wasserwesen (NAW) im DIN - Water quality - Determination of the inhibitory effect of water samples on the light emission of Vibrio fischeri (luminescent bacteria test) Part 2: Method using liquid-dried bacteria ISO 11348-2:2007

Pavlić, Ž., Vidaković-Cifrek, Ž., & Puntarić, D. (2005) Toxicity of surfactants to green microalgae Pseudokirchneriella subcapitata and Scenedesmus subspicatus and to marine diatoms Phaeodactylum tricornutum and Skeletonema costatum. Chemosphere, 61:1061-1068

Ratte 2020: Statistical Evaluation of a round robin test as part of the EU Project Sullied Sediments: Determination of the toxic effect of sediment samples using the Algae Growth Inhibition Test (based on DIN 38-412 L33 and DIN EN ISO 8692)1 and the Bacteria Luminescence Test (DIN EN ISO 11-348). Unpublished report.

Regoli, F., d'Errico, G., Nardi, A., Mezzelani, M., Fattorini, D., Benedetti, M., Di Carlo, M., Pellegrini, D., Gorbi, S. (2019): Application of a Weight of Evidence Approach for Monitoring Complex Environmental Scenarios: the Case-Study of Off-Shore Platforms. Frontiers in Marine Science. 6: Article 377.

Rojíčková R, Maršálek B (1999) Selection and sensitivity comparisons of algal species for toxicity testing. Chemosphere, 38:3329-3338

Swartz RC, DeBen WA, Cole FA (1979) A Bioassay for the Toxicity of Sediment to Marine Macrobenthos. Journal WPCF (Water Pollution Control Federation), 51:944-950

Teuchies, J., WP3 partners from Sullied sediments project (2017): Sullied sediments – WP3 Sediment assessment Sampling SOP.

USEPA-USACOE (1998) Evaluation of dredged material proposed for discharge in waters of the U.S.-Inland testing manual. EPA-823-B-98-004. US Environmental Protection Agency, Office of Water, and US Army Corps of Engineers, Washington, DC

Wetzel MA, Wahrendorf D-S, von der Ohe PC (2013) Sediment pollution in the Elbe estuary and its potential toxicity at different trophic levels. Science of The Total Environment, 449:199-207

Zovko, M., Vidaković-Cifrek, Ž., Cvetković, Ž., Bošnir, J., & Šikić, S. (2015). Assessment of acrylamide toxicity using a battery of standardised bioassays. Arhiv Za Higijenu Rada i Toksikologiju, 66(4): 315–321. https://doi.org/10.1515/aiht-2015-66-2715

# Acknowledgements

‘Sullied Sediments’ has been co-funded by the European Regional Development Fund through the Interreg VB North Sea Region Programme with a grant of 2.043.413 euros with equivalent match funding from the partners involved. The project partnership includes public, private and third sector organisations based in the United Kingdom, Germany, Belgium and the Netherlands

The present project was funded by European Regional Development Fund. Concerning Chapter 2 discussions with Nicole Schlichting (IDN) how to conduct the round robin test were greatly appreciated. Chemical and physical properties of sediments 1-3 and 5 and 7 were provided by Sullied Sediments Partners IDN (Katharina Heitmann) and SOCOTEC (Paul Walker).

.

Interreg
North Sea Region
European Regional Development Fund

EUROPEAN UNION

# Partners

**The Sullied Sediments project partnership comprises 13 project beneficiaries:**

*Canal and River Trust (UK)*

*East Riding of Yorkshire Council (UK)*

*Ecossa (Germany)*

*Hamburg Port Authority (Germany)*

*Hamburg University of Applied Sciences (Germany)*

*Institut Dr. Nowak* GmbH & Co. KG *(Germany)*

*Openbare Vlaamse Afvalstoffenmaatschappij (Belgium)*

*Radboud University (The Netherlands)*

*Socotec UK Ltd (UK)*

*University of Antwerp (Belgium)*

*University of Hull (UK)*

*University of Leeds (UK)*

*Vlaamse Milieumaatschappij (Belgium)*

**The partnership also receives expert advice from 12 strategic partners who form our Advisory Group:**

*East and North Yorkshire Waterways Partnership (UK)*

*Elbe Habitat Foundation (Germany)*

*Environment Agency (UK)*

*Federal Institute of Hydrology (Germany)*

*Foundation for Applied Water Research (Europe)*

*Hamburg Ministry of the Environment and Energy (Germany)*

*Northumbrian Water (UK)*

*River Hull Board (UK)*

*Sediment European Network Steering Group (European)*

*Thames Water (UK)*

*Vlakwa (water research consultancy) (Belgium)*

*Yorkshire Water (UK)*

# Contact us

**For more information about these publications, please contact:**

Visit our webspace: www.northsearegion.eu/sullied-sediments

Follow us on Twitter: @SulliedSediment

Join our blog: www.sulliedsediments.wordpress.com/

Join our blog: www.sulliedsediments.wordpress.com/

# Appendices/Supporting info

**Annex 1: Exemplary calculation of pore water percentage of elutriates for volumetric and gravimetric methods for G1 (highest concentration):**

Example 1: Sediment A with 23 % dry weight; 77 % water

Example 2: Sediment B with 34 % dry weight; 66 % water

**Gravimetric method (Lab 2, IDN):**

Example 1: Sediments with a water content of more than 75 % are elutriated by their own water, no additional water is added:

Elutriate after centrifugation contains 100 % "porewater", and after addition of media and algae: 80 % in the G1 (highest concentration).

Example 2: The sample is adjusted to a liquid/sediment (dw) ratio of 3:1, whereby the "liquid" will contain the porewater. 125 g dw of Sediment B will have a wet weight of 367.65 g, containing 242.65 g porewater. As total aqueous volume for elutriation is 375 g (3 times 125 g), the porewater is diluted with (375 g -242.65 g) 132.35 g of water. The percentage in the elutriate after centrifugation will thus contain 64.7 %, and after addition of media and algae 51.7 % in the G1 (highest concentration).

**Volumetric method (Lab 1, HAW):**

Disregarding the porewater content, 100 ml of Sediment are resuspended in 400 ml of water.

In order to calculate the porewater content of the elutriate after centrifugation as percentage, the volume porewater divided by the sum (porewater + 400 ml), times 100.

Example 1: 77*100/(77+400)=16,14 % in the elutriate after centrifugation; 12.91 % in the G1

Example 2: 66 g*100/(66+400)=14.16 % in the elutriate after centrifugation; 11.33 % in the G1

In the examples above, the higher density of the sediment was not taken into account. As no sediment density had been determined, a comparison was calculated assuming an average sediment density of 1.7 g/cm$^3$ for sediments from the Elbe estuary[2] to calculate the sediment mass from volume.

Example 1: 77%*170 g=130.9 g Porewater;

Porewater percentage in the elutriate after centrifugation: 130.9 g*100/(130.9+400)=24.66 %; 19.7 % in G1

Example 2: 66 %*170 g=112.2 g Porewater

Porewater percentage in the elutriate after centrifugation: 112.2 g*100/(112.2+400)=21.9 %; 17.5 % in G1

For these sediments, a density of 1.7 g/cm$^3$ is certainly too high. Nevertheless, the difference in porewater percentage is below 10 %

*Table 9: Overview of porewater content of elutriates in % in G1 calculated for 2 exemplary sediments*

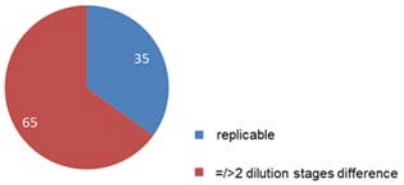
| | Volumetric method | | Gravimetric method |
|---|---|---|---|
| | Not corrected for sediment density | Assuming sediment density of 1.7 g/cm$^3$ | |
| Example 1 | 12.92 | 19.7 | 80 |
| Example 2 | 11.33 | 17.5 | 51.7 |

[2] Based on calculations of Vollmer et al. (2014) for the middle Elbe (doi: 10.5675/Kliwas_67/2014_Binnenelbe)

## Annex 2:

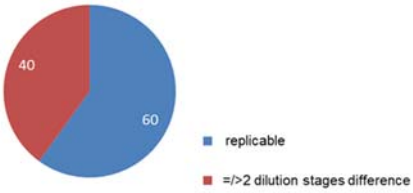
| FWAT | | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 |
|---|---|---|---|---|---|---|---|---|
| n=14 | pT | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Lab2 | 0 | 1 | | 3 | | 1 | 1 | |
| Lab2 | 1 | | | | 1 | | | |
| Lab2 | 2 | | 1 | | 2 | | 1 | |
| Lab2 | 3 | | | | | 1 | 1 | |
| Lab2 | 4 | | | 1 | | | | |
| Lab2 | 5 | | | | | | | |
| Lab2 | 6 | | | | | | | |

A




| FWAT | | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 | Lab1 |
|---|---|---|---|---|---|---|---|---|
| n=23 | pT | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Lab2 | 0 | 3 | 5 | 3 | 2 | | | |
| Lab2 | 1 | | 1 | 2 | | | | |
| Lab2 | 2 | | | | | | | |
| Lab2 | 3 | | | | | 1 | | |
| Lab2 | 4 | | | 1 | 1 | | 1 | 2 |
| Lab2 | 5 | | | | 1 | | | |
| Lab2 | 6 | | | | | | | |

B




*Annex 1: Interlaboratory comparison of AGI results. A: Test was run in 2012, B: Test was run in 2015*

## Annex 3

Ratte (2020): Report on "Statistical Evaluation of a round robin test as part of the EU Project Sullied Sediments: Determination of the toxic effect of sediment samples using the Algae Growth Inhibition Test (based on DIN 38-412 L33 and DIN EN ISO 8692) and the Bacteria Luminescence Test (DIN EN ISO 11-348) (unpublished).

# Statistical Evaluation of a round robin test as part of the EU Project Sullied Sediments: Determination of the toxic effect of sediment samples using the Algae Growth Inhibition Test (based on DIN 38-412 L33 and DIN EN ISO 8692)[1] and the Bacteria Luminescence Test (DIN EN ISO 11-348)

---


| | |
|---|---|
| Organization of the Round robin Test | Dr. Annette Kramer<br>Hamburg Port Authority, D-20457 Hamburg |
| Financial Support for Evaluation and Report | Interreg North Sea Region Program ERDF<br>Hamburg Port Authority, D-20457 Hamburg |
| Statistical Evaluation and Report | Dr. Monika Ratte, ToxRat Solutions GmbH & Co. KG, DE |


**Final Report: April-20-2020**

---

[1] Flöter, Carolin (2009): Basisdokument zur Probenvorbereitung von limnischen Sedimenten für die Durchführung von limnischen Biotestverfahren im Auftrag der Hamburg Port Authority, Stand 2009_03_17 [1]

## Summary

The EU Project Sullied Sediments aims at collecting data and testing new tools to better assess, treat and prevent contamination from certain pollutants that can be found in sediments. Through Work Package 3 - Sediment Assessment, a round robin test with six participating laboratories (L01 – L06) was performed in October and November 2018, using the Algae Growth Inhibition Test (AGI) and the Luminescence Bacteria Test (LBT) as test systems to assess sediment toxicity. The objectives of the round robin test and the statistical evaluation were:

- To characterize the variability of the ecotoxicological endpoints "Lowest Ineffective Dilution" (LID), which is defined as the lowest dilution causing an inhibition lower than 20%, and the "percentage of inhibition in D1", i.e. in the undiluted sample.
- To investigate whether the AGI and LBT are suitable to assess toxicity of different sediments.
- To investigate whether the "percentage of inhibition in D1" is correlated to the LID.
- To characterize the performance of the participating laboratories.

Samples from seven natural sediments (Sed 1 - Sed 7) were sent to the participants. For Sed 1, Sed 2 and Sed 3, two samples were provided each, without being labeled as duplicate samples. In total, each laboratory received 10 sediment samples. For every sediment, eluate and pore water samples were prepared by the individual laboratories and subsequently tested with AGI and LBT, each.
The total number of valid tests for Sed 4-7 was lower or equal to 6, with sometimes even going down to 3. For Sed 1-3, the total number of valid tests was between 7 and 12, based on measurements of 4 to 6 laboratories. The present evaluation was performed twice: one time based only on valid data and a second time based on all data.

Means and reproducibilities were calculated according to DIN ISO ISO 5725-2. Additionally, the normalized width (NW) of an obtained mean was calculated as ratio between the mean and the range of its 95% confidence interval. A measure for laboratory performance was calculated adapting the principle of z-scores according to ISO 38402-45. Special modifications were applied to account for the special characteristics of the present database. To keep it simple, the measure for variability is nevertheless named reproducibility. To prevent miss-interpretation of the calculated scores, they are called pseudo-z-scores.
Results based on valid data were sometimes slightly different from those based on all data, but basically in the same range. The following conclusions therefore focus on the results from valid data. In view of the very limited data base, any results should be regarded as conditionally.

**Variability of the measurement variables**
Repeatability of LID was assessed by the duplicate samples. In the AGI, 26 out of 30 duplicate samples showed identical LIDs (20) or LIDs which differ not more than one dilution step (6). In the LBT, all 36 duplicate samples showed identical LIDs (32) or LIDs which differ not more than one dilution step (4). Reproducibility, expressed as coefficient of variation of total variability (CV% sR), of LID in the AGI in most cases exceeded 100% and NWs were up to 10. In the LBT, some mean LIDs also showed CV% sR up to 100% and NWs up to 10, but generally the LBT provided better reproducibility than the AGI.

Repeatability CV% sr for "%inhibition in D1" in most cases of AGI and LBT was lower than 50% or even lower than 20%.

The reproducibility CV% sR of "%inhibition in D1" in most cases were around 80% or even higher, with the AGI showing the higher percentages. In most cases, values for NW between 2 and 3 were obtained, indicating a generally better reliability of "%inhibition in D1" compared to LID.

**Suitability of the AGI and LBT to assess toxicity of different sediments**

The toxicity of each sediment was assessed using the mean of all laboratory specific LID values obtained for this sediment. If the results from eluate and pore water were different, the higher mean LID (and thus toxicity) was taken.

Based on mean LIDs, Sed 1 and 4 were identified as the most toxic sediments in the round robin test both by AGI and LBT. Nevertheless, the broad 95% confidence ranges (CI) of the mean LIDs indicate their high uncertainty (sed 1: mean LID = 5.9 (AGI, 95% CI 0.8 - 42.7) and 4.5 (LBT, CI = 1.5 - 13.9); sed 4: mean LID = 11.3 (AGI, 95% CI 1.7 - 76.4) and 3.5 (LBT, CI 1.0 - 12.5).

Sed 3, 5, 6 and 7 were consistently found to be of low mean toxicity or non-toxic in both tests (mean LID = 1.0 – 2.7 (AGI) and mean LID = 1.1 – 2.5 (LBT), with the corresponding 95% CIs between 0.2 - 26 (AGI) and 0.7 - 4.6 (LBT)).

For sed 2, there were no consistent means: while the mean LID from LBT was 1.4, indicating very low toxicity (95% CI 0.7 – 2.7), the mean LID from AGI indicates moderate toxicity (mean LID = 6.5, 95% CI = 0.8 – 51.7), i.e. the obtained mean LID of the AGI shows high uncertainty.

When comparing the results of eluate and porewater samples, in the AGI, three sediments (sed 1, 2 and 4), revealed markedly different mean LIDs and thus toxicity in eluate and pore water samples, with two times eluate showing the higher mean toxicity (sed 2 and 4) and one time pore water showing the higher mean toxicity (sed 1). In four sediments (sed 3,5,6 and 7), the mean LIDs measured in eluate and pore water samples were quite similar in each case. In the LBT, eluate and pore water samples revealed very similar mean LIDs, except for sed 1 where the mean LID found in eluate was about twice that of pore water.

Regarding "%inhibition in D1", in the AGI, the means of eluate and pore water agreed quite well for sediment 4, 5, 6 and 7, whereas for sediment 1, 2 and 3, generally the eluate samples showed markedly higher "%inhibition in D1". In the LBT, eluate and pore water samples showed quite similar results with a tendency to slightly higher mean "%inhibitions in D1" in pore water.

To sum up, in terms of mean LIDs, both AGI and LBT showed consistent (but nevertheless highly variable) results regarding the general toxicity level measured, except for one sediment (sed 2). Thereby, the overall level of mean LIDs detected by the LBT was basically lower compared to those of the AGI or equal in case of anyway low toxicity detected by the AGI. The precision of the obtained mean LIDs is very low and thus the reliability of of a single result from a single laboratory also is found to be very poor. Thereby, the results from LBT were less variable than those from AGI. This might be due to the overall lower sensitivity of the LBT and / or to lower degree of interferences due to cofounding factors.

**Relationship between the LID and the percentage of inhibition in D1**

Regarding the overall means per sediment, the mean LID for a certain sample and the corresponding mean "%inhibition in D1", revealed quite consistent results both in AGI and in LBT, i.e. the higher the "%inhibition in D1", the higher the LID. When analyzing the individual results per laboratory, however, a different picture was obtained. Whereas for the LBT, there was a clear correlation between both variables ($R^2$ = 0.64), no clear relationship between "%inhibition in D1" and LID was found for the AGI ($R^2$ = 0.12).

**Performance of the participating laboratories**

The individual laboratory performance was assessed by the ability to perform a valid test and the standardized deviation of the single laboratory results from the corresponding overall mean result, assessed by pseudo-z-sores. For the AGI, 14 out 36 controls failed to meet the validity criteria (39%). All invalid controls were derived by L05 (4 out of 10 controls invalid, i.e. 40% of the controls run by L05) and L06 (10 out of 14 controls invalid, i.e. 71% of the controls run by L06). For the LBT, 9 out of 62 controls were found to be invalid (15%). L06 accounted for most of the invalid controls (7 out of 20, 35% of the controls run by L06).

L03 tends to detect higher values for toxicity measures in the LBT compared to other laboratories (mean pseudo z-score LID = 0.972, "%inhibition in D1" = 1.371). L04 tends to lower values for toxicity in the LBT (mean pseudo z-score LID = - 0.691; "%inhibition in D1" = -1.441) and slightly higher values for toxicity in the AGI (mean pseudo-z-score LID = 0.606).

**Overall conclusion**

Overall, it should be clearly stated, that the data base of the present round robin test, especially the number of participating laboratories, is too poor to derive robust statistical results. No serious conclusions can be derived except the fact that the LBT seems to be less sensitive then the AGI and that LIDs obtained in AGI and LBT tests are highly variable, with the LBT showing slightly better reproducibility than the AGI. Moreover, the high number of invalid tests in single laboratories points to possible shortcomings with test performance.

The data set and results presented in this report should be taken as a preliminary interlaboratory comparison providing a first orientation and helpful hints in order to prepare a more comprehensive round robin test.

# Table of Contents

# 1. Introduction

The EU Project Sullied Sediments aims at collecting data and testing new tools to better assess, treat and prevent contamination from certain pollutants that can be found in the sediments in our waterways. Through Work Package (WP) 3 - Sediment Assessment, this project intends to provide better tools for sediment assessment in order to enable better risk assessment and reduce economic costs.
The project was started in January 2017. During the course of WP3, a round robin test with six participating laboratories was performed in October and November 2018, investigating the suitability of two biotests for assessing sediment toxicity: The Algae Growth Inhibition Test (AGI) based on DIN 38-412 L33 and DIN EN ISO 8692[2] and the Bacteria Luminescence Test (LBT, DIN EN ISO 11-348-2). Both tests are designed to measure the ecotoxicological potential of sediments by exposing the test organisms to different dilutions of pore water and eluate of the tested sediment. Instructions for the preparation of pore waters and eluates were given in an extra document[3] provided to each laboratory. The principle of the tests is to measure certain biological parameters such as biomass production (AGI) or luminescence (LBT) in different dilutions of the samples and to calculate the corresponding percentages of inhibition compared to a negative control. The toxicological endpoint is the lowest ineffective dilution (LID), which is defined as the lowest dilution causing an inhibition lower than 20%[4]. Within the project “Sullied Sediments” it is argued that the percentage of inhibition in D1, i.e. in the undiluted sample, maybe of higher ecological relevance and thus be more relevant for ecotoxicological risk assessment than the overall LID, since in reality, organisms are exposed to original substance concentrations rather than to dilutions. Therefore, the present statistical evaluation of the tests focuses on both the obtained LIDs and the measured percentages of inhibition in D1.

The objectives of the round robin test and the statistical evaluation were:

- To characterize the variability of the ecotoxicological endpoints LID and “% inhibition in D1”,
- to investigate whether the AGI and LBT are suitable to assess toxicity of different sediments,
- to investigate whether the “% inhibition in D1” is correlated to the LID, and
- to characterize the performance of the participating laboratories.

[2] Test was performed according to „Basisdokument zum Grünalgentest mit Desmodesmus subspicatus“ “ which is mainly based on DIN 38412-L33 [1]
[3] Flöter, Carolin (2009): Basisdokument zur Probenvorbereitung von limnischen Sedimenten für die Durchführung von limnischen Biotestverfahren im Auftrag der Hamburg Port Authority, Stand 2009_03_17 [1a]
[4] For details see BfG-Merkblatt "Ökotoxikologische Baggergutuntersuchung" [2]

# 2. Data base, Validity and Methods for Statistical Evaluation

## 2.1. Participants

Six laboratories participated in the round robin test (see Table 1), labeled by a random laboratory code between L01 and L06 in this report.

**Table 1: Participating laboratories in alphabetical order**

| Laboratory |
|---|
| ECT Oekotoxikologie GmbH |
| Eurofins Umwelt Ost GmbH |
| HAW Hamburg University of Applied Sciences (working group Applied Aquatic Toxicology) |
| Hydrotox GmbH |
| Institut Dr. Nowak |
| Noack Laboratorien GmbH |

## 2.2. Validity and data base for statistical evaluations

The experiments were performed in October and November 2018. Samples from seven natural sediments (Sed 1 - Sed 7) were sent to the participating laboratories by the round robin test organizer. For Sed 1, Sed 2 and Sed 3, two samples were provided per sediment, without being labeled as duplicate sample. Hence, each laboratory received 10 sediment samples. Sediment samples were taken following a synchronized sampling protocol (Teuchies et al., 2017 [6]). Sediments from one source were homogenized in a standardized way and then divided into sub-samples that were sent to the laboratories. For every sediment sample, eluate and pore water samples were prepared by the individual laboratory and subsequently tested with AGI and LBT, each. For sediments 1 - 3, results from two parallel samples per laboratory and biotest were available, whereas for sediments 4-7, results from one sample per laboratory and biotest were obtained.

The validity of the biotests results was assessed by the performing laboratories and rechecked by the round robin test organizer according to the validity criteria prescribed in the corresponding test guidelines or SOP. Thereby, the following exceptions from the formal validity criteria were applied:

- Algae Growth Inhibition test, L01, Sed 6, Eluate: Blank values were slightly enhanced; however, the laboratory reported no evidence for existence of autochthonous algae; therefore, the test was rated as valid.
- Algae Growth Inhibition test, L05: Eluate of Sed 1 (sample 2), Sed 2 (sample 2) and Sed 3 (sample 1+2): the reference sample used for these 4 tests showed an inhibition of 82,5%, whereas a maximum of 80% is allowed. The performing laboratory identified the samples as valid, which was accepted in view of the small violation of the limit.

Apart from formal validity criteria, there was a special issue with samples of L06 in the LBT:

- in most Bacteria Luminescence tests of L06, the storage time of samples (time between sediment sampling and test start) exceeded 30 days. Provided that the tests were valid with respect of all other validity criteria, the results were nevertheless used for statistical evaluation. This applies to the following samples: Sed 1 and Sed 2, Eluate 1 + 2; Sed 3, Eluate 1+2 and Porewater 1+2; Sed 4 and Sed 5, Eluate; Sed. 7 Porewater.

The following factors reduced the data base available for statistical analysis:

- Several laboratories used one control for multiple tests. As a consequence, if this control failed to meet the validity criteria, several tests had to be classified as invalid.
- L03 didn´t perform the AGI at all.
- L05 didn´t test the treatment D1 (i.e. undiluted sample) in the LBT, hence no data for percentage of inhibition is available.

The following special issues need to be considered:

- Some laboratories did not test the prescribed dilutions, but fewer ones or other ones. Generally, dilution series for LBT should have been proceeded until inhibition is below 20%. This was not done by L03 in every single test for sed 1, 4, 7. Further, L05 did not test the inhibition values for the D1 (see above) Additionally, in the AGI, the critical inhibition of 20% in some tests was still exceeded in the highest dilution (D32). As a consequence, in these cases, the LID could only be judged as $\underline{<}$ x or $\underline{>}$ x. To enable statistical evaluation, these results were set to x.
- Contrary to the requirements given by the test organizer for the LBT, Lab 03 did not apply the color correction for the inhibition values. Therefore, the corrections were calculated for each dilution stage based on the results from Lab 01 who provided color corrected values and not corrected values.
- The dilution series is based on the dilution factor 2, i.e. the LID can only be a discrete number out of 6 values (namely, 1,2,4,8,16 or 32). In most cases, only 2-4 different results are obtained for a single sediment.
- LID values are log normal distributed and thus had to be transformed for statistical evaluation to ensure normal distribution. The results then had to be retransformed.
- In several tests, stimulations were observed in D1, i.e. negative percentages for inhibition were reported. As only inhibitory effects are of interest and since the statistical evaluations require

only positive numbers, (see section 2.3), all stimulations were generally set to zero percent inhibition.

Table 2 and Table 3 show the data base used for the present statistical evaluation, i.e. with the modifications performed to original data as explained above. The original results are provided in Annex Table 1 and 2.

Any evaluations were performed separately for results from eluate samples and from pore water samples. If the mean LIDs from eluate and pore water were different, the higher LID was taken to assess the toxicity of the corresponding sediment.

**Table 2:** Results of the Algae Growth Inhibition test; data base as used for statistical evaluation. PW = Pore water, EL = Eluate; LID = Lowest inhibitory dilution; LIDs ≤ x or ≥ x set to x; Negative % inhibitions (i.e. stimulations) set to zero. Original data available in Annex Table 1.
Sediment 1,2 and 3: two different samples of the same sediment were tested.
Red: test invalid.

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| PW | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 2 | 1 | 1 | 8 | 1 | 1 | 2 |
| | 2 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 4 | 1 | 1 | 1 |
| | 32 | 1 | 1 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 16 | 16 | 16 | 2 | 1 | 4 | 1 |
| | 16 | 16 | 16 | | | | |
| L5 | 2 | 1 | 4 | 4 | 1 | 2 | 1 |
| | 2 | 4 | 2 | | | | |
| L6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | | | | |

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| PW | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 42,3 | 0,0 | 0,0 | 85,5 | 0,0 | 0,0 | 57,7 |
| | 29,9 | 0,0 | 0,0 | | | | |
| L2 | 27,0 | 0,0 | 0,0 | 91,0 | 0,0 | 0,0 | 0,0 |
| | 47,0 | 0,0 | 0,0 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 0,0 | 0,0 | 0,0 | 70,0 | 0,0 | 0,0 | 0,0 |
| | 0,0 | 0,0 | 0,0 | | | | |
| L5 | 20,0 | 0,0 | 32,3 | 64,2 | 0,0 | 52,2 | 0,0 |
| | 41,3 | 29,5 | 27,1 | | | | |
| L6 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 |
| | 0,0 | 16,8 | 0,0 | | | | |

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| EL | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 4 | 2 | 1 | 4 | 1 | 1 | 2 |
| | 2 | 1 | 2 | | | | |
| L2 | 2 | 4 | 1 | 32 | 4 | 8 | 1 |
| | 4 | 32 | 1 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 1 | 16 | 32 | 32 | 4 | 1 | 4 |
| | 1 | 16 | 8 | | | | |
| L5 | 2 | 8 | 1 | 4 | 1 | 2 | 1 |
| | 4 | 8 | 1 | | | | |
| L6 | 1 | 8 | 1 | 8 | 1 | 1 | 1 |
| | 1 | 8 | 1 | | | | |

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| EL | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 62,0 | 28,1 | 12,1 | 93,1 | 0,0 | 0,0 | 49,9 |
| | 44,4 | 11,6 | 23,7 | | | | |
| L2 | 80,0 | 89,0 | 0,0 | 66,0 | 0,0 | 4,0 | 10,0 |
| | 74,0 | 81,0 | 6,0 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 0,0 | 64,0 | 93,0 | 41,0 | 49,0 | 0,0 | 47,0 |
| | 0,0 | 50,0 | 90,0 | | | | |
| L5 | 49,1 | 0,0 | 11,9 | 79,1 | 0,0 | 22,2 | 0,0 |
| | 48,8 | 25,7 | 0,0 | | | | |
| L6 | 30,5 | 0,0 | 0,0 | 25,5 | 0,0 | 0,0 | 0,0 |
| | 0,0 | 0,0 | 0,0 | | | | |

**Table 3:** Results of the Bacteria Luminescence test; data base as used for statistical evaluation. PW = Pore water, EL = Eluate; LID = Lowest inhibitory dilution; LIDs ≤ x or ≥ x set to x; Negative % inhibitions (i.e. stimulations) set to zero. Original data available in Annex Table 2.
Sediment 1,2 and 3: two different samples of the same sediment were tested.
Red: test invalid.

| Bacteria Luminescence Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| PW | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 1 | 1 | 1 | 4 | 4 | 1 | 1 |
| | 1 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 2 | 2 | 1 | 2 |
| | 4 | 1 | 1 | | | | |
| L3 | 8 | 4 | 2 | 16 | 4 | 2 | 8 |
| | 8 | 2 | 2 | | | | |
| L4 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| | 1 | 1 | 1 | | | | |
| L5 | 2 | 2 | 2 | 4 | 4 | 2 | 2 |
| | 2 | 2 | 2 | | | | |
| L6 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 2 | | | | |

| Bacteria Luminescence Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| PW | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 16,2 | 3,5 | 4,6 | 27,0 | 32,9 | 0,0 | 11,4 |
| | 15,2 | 1,8 | 8,3 | | | | |
| L2 | 42,3 | 17,6 | 13,9 | 32,0 | 27,7 | 12,1 | 28,4 |
| | 47,6 | 17,7 | 15,2 | | | | |
| L3 | 57,1 | 31,8 | 31,2 | 87,4 | 72,2 | 41,8 | 62,9 |
| | 58,7 | 33,1 | 28,3 | | | | |
| L4 | 0,0 | 0,0 | 0,0 | 0,0 | 0,0 | 57,7 | 10,9 |
| | 0,0 | 0,0 | 0,0 | | | | |
| L5 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| | n.d. | n.d. | n.d. | | | | |
| L6 | 31,0 | 27,5 | 23,4 | 4,6 | 1,3 | 0,0 | 18,8 |
| | 29,9 | 18,0 | 10,3 | | | | |

| Bacteria Luminescence Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| EL | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 4 | 1 | 1 | 4 | 1 | 1 | 1 |
| | 4 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 2 | 1 | 1 | 1 |
| | 4 | 1 | 1 | | | | |
| L3 | 16 | 1 | 1 | 8 | 4 | 2 | 2 |
| | 16 | 1 | 1 | | | | |
| L4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | | | | |
| L5 | 16 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 8 | 2 | 2 | | | | |
| L6 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| | 2 | 1 | 1 | | | | |

| Bacteria Luminescence Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| EL | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 36,2 | 1,8 | 0 | 33,2 | 0,0 | 0,0 | 1,1 |
| | 48,7 | 0,0 | 0 | | | | |
| L2 | 43,6 | 15,1 | 2,9 | 48,5 | 7,6 | 12,9 | 5,4 |
| | 43,4 | 5,5 | 6,8 | | | | |
| L3 | 76,8 | 15,4 | 15,7 | 66,6 | 57,6 | 46,5 | 50,2 |
| | 86,8 | 16,4 | 10,8 | | | | |
| L4 | 0,0 | 0,0 | 0 | 0,0 | 0,0 | 0,0 | 0,0 |
| | 0,0 | 0,0 | 0 | | | | |
| L5 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| | n.d. | n.d. | n.d. | | | | |
| L6 | 0,0 | 3,5 | 0,0 | 43,1 | 3,9 | 0,9 | 1,4 |
| | 36,4 | 0,0 | 0,0 | | | | |

Since some laboratories used the same control for multiple biotests, the number of biotests performed differs from the number of controls used. Thus, the number of valid controls differs from the number of valid tests. Table 4 and Table 5 show the controls used by the different laboratories in the different tests, their validity and the resulting number of valid tests per laboratory and per sediment.

The total number of valid tests per method for Sed 4-7 is lower or equal to 6, each, with sometimes even going down to 3. For Sed 1-3, the total number of valid tests is between 7 and 12, based on measurements of 4 to 6 laboratories. According to the round robin test organizer, “the number of invalid tests showed the susceptibility of the methods and reflect the daily routine in handling ecotoxicological test results”. Therefore, the test organizer decided to perform the statistical evaluation twice: once based on only valid data and a second time based on all data.

**Table 4:** Controls used by the different laboratories in the different Algae Growth Inhibition tests and resulting number of valid tests. PW = pore water, EL = eluate; Red: test invalid.
Reading Example: "L05 performed 20 Algae Growth Inhibition tests in total, using 10 different controls. C1, C3, C5 and C6 were found to be invalid, hence all tests based on these controls were judged as invalid (namely 8 tests)."

| Algae Growth Inhibition Test | | | | | | | | number of valid tests |
|---|---|---|---|---|---|---|---|---|
| PW | Controls used (red = invalid) | | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 | |
| L1 | C1 | C1 | C1 | C2 | C2 | C2 | C3 | 10 |
| | C1 | C1 | C1 | | | | | |
| L2 | C1 | C2 | C2 | C2 | C2 | C2 | C2 | 10 |
| | C2 | C2 | C1 | | | | | |
| L3 | - | - | - | - | - | - | - | - |
| | - | - | - | | | | | |
| L4 | C1 | C1 | C1 | C2 | C2 | C2 | C2 | 10 |
| | C1 | C1 | C1 | | | | | |
| L5 | C1 | C1 | C2 | C4 | C4 | C5 | C5 | 4 |
| | C2 | C3 | C3 | | | | | |
| L6 | C1 | C2 | C4 | C6 | C7 | C7 | C8 | 2 |
| | C1 | C3 | C5 | | | | | |
| **no of valid tests** | **7** | **8** | **7** | **4** | **4** | **3** | **3** | |

| Algae Growth Inhibition Test | | | | | | | | number of valid tests |
|---|---|---|---|---|---|---|---|---|
| EL | Controls used (red = invalid) | | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 | |
| L1 | C4 | C4 | C4 | C5 | C5 | C5 | C6 | 10 |
| | C4 | C4 | C4 | | | | | |
| L2 | C1 | C1 | C1 | C2 | C2 | C2 | C2 | 10 |
| | C1 | C1 | C2 | | | | | |
| L3 | - | - | - | - | - | - | - | - |
| | - | - | - | | | | | |
| L4 | C3 | C3 | C3 | C4 | C4 | C4 | C4 | 10 |
| | C3 | C3 | C3 | | | | | |
| L5 | C6 | C6 | C7 | C9 | C9 | C10 | C10 | 8 |
| | C7 | C8 | C8 | | | | | |
| L6 | C9 | C10 | C11 | C13 | C14 | C13 | C14 | 2 |
| | C9 | C10 | C12 | | | | | |
| **no of valid tests** | **7** | **7** | **8** | **4** | **5** | **4** | **5** | |

**Table 5:** Controls used by the different laboratories in the different Bacteria Luminescence tests. PW = pore water, EL = eluate; Red: test invalid.
Reading Example: "L03 performed 20 Bacteria Luminescence tests in total, using 4 different controls. All tests with PW samples of Sed 4, 5, 6 and 7 were related to the same control C2. Since this control was found to be valid, the corresponding tests were valid, too."

| Bacteria Luminescence Test | | | | | | | | number of valid tests |
|---|---|---|---|---|---|---|---|---|
| PW | Controls used (red = invalid) | | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 | |
| L1 | C1 | C1 | C1 | C2 | C2 | C2 | C3 | 10 |
| | C1 | C1 | C1 | | | | | |
| L2 | C1 | C1 | C1 | C2 | C2 | C3 | C2 | 10 |
| | C1 | C1 | C1 | | | | | |
| L3 | C1 | C1 | C1 | C2 | C2 | C2 | C2 | 10 |
| | C1 | C1 | C1 | | | | | |
| L4 | C1 | C2 | C2 | C3 | C3 | C3 | C3 | 10 |
| | C1 | C1 | C1 | | | | | |
| L5 | C1 | C3 | C5 | C7 | C8 | C9 | C10 | 9 |
| | C2 | C4 | C6 | | | | | |
| L6 | C1 | C3 | C5 | C7 | C8 | C9 | C10 | 5 |
| | C2 | C4 | C6 | | | | | |
| **no of valid tests** | **12** | **10** | **11** | **5** | **6** | **4** | **6** | |

| Bacteria Luminescence Test | | | | | | | | number of valid tests |
|---|---|---|---|---|---|---|---|---|
| EL | Controls used (red = invalid) | | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 | |
| L1 | C4 | C5 | C5 | C6 | C6 | C6 | C7 | 10 |
| | C5 | C5 | C4 | | | | | |
| L2 | C4 | C1 | C5 | C2 | C3 | C3 | C6 | 10 |
| | C4 | C5 | C5 | | | | | |
| L3 | C1 | C3 | C3 | C2 | C4 | C2 | C4 | 10 |
| | C3 | C3 | C3 | | | | | |
| L4 | C4 | C4 | C4 | C5 | C5 | C5 | C5 | 9 |
| | C4* | C4 | C4 | | | | | |
| L5 | C11 | C13 | C15 | C17 | C18 | C19 | C20 | 10 |
| | C12 | C14 | C16 | | | | | |
| L6 | C11 | C13 | C15 | C17 | C18 | C19 | C20 | 8 |
| | C12 | C14 | C16 | | | | | |
| **no of valid tests** | **11** | **12** | **11** | **6** | **5** | **6** | **6** | |
| | * test invalid because of too high variability in a treatment | | | | | | | |

## 2.3. Statistical Analysis

Precision of the round robin tested biotests was analyzed according to DIN ISO 5725-2 by means of reproducibility of the results. The quality of the results in terms of "correctness" and thus the performance of the participating laboratories was assessed by means of the standardized deviation of the measured values from the corresponding mean value, based on the z-score concept described in DIN 38402-45 [4]. In both cases, individual adaptations were performed due to the specific character of biotests in general and due to specific characteristics of the data base.

According to DIN ISO 5725-1, "it is common to choose a value between 8 and 12" for the number of participating laboratories and "estimates of the repeatability and reproducibility standard deviations could differ substantially from their true values if only a small number (p ~ 5) of laboratories take part in a precision experiment". It should thus be clearly stated, that the data base available for statistical evaluation is too poor to derive robust statistical results and that the results presented in this report should be taken just as ranges and tendencies providing a first orientation.

### 2.3.1. Repeatability and Reproducibility

The precision of the round robin test results was assessed according to DIN ISO 5725-2, with certain modifications due to the special characteristics of the data.
The total variance ($s^2R$ , reproducibility variance) is based on the variance within a single laboratory, i.e. the intra-laboratory variance ($s^2r$, also called repeatability variance), and the variance between laboratories, i.e. the inter-laboratory variance ($s^2L$ ), see Eq1).

$$s_R^2 = s_L^2 + s_r^2 \qquad (1)$$

The corresponding standard deviations then can be calculated as square roots. In the following text, the terms "reproducibility" and "repeatability" are also used for the total standard deviation sR and sr, respectively. "Coefficient of variation (CV% sR or CV% sr )" means the reproducibility sR or the repeatability sr expressed as percentage. A high CV% sR means a high variability and thus a poor reproducibility.
According to DIN ISO 5725-2, assessment of repeatability requires that a sample is measured several times. In the present round robin test, the duplicate samples tested for sediment 1, 2 and 3 by every laboratory were used to determine intra-laboratory variance. The duplicate samples were taken as sub-samples from the same source sediment that was homogenized in a synchronous way (see chapter 2.2.). L01, L02 and L03 used the same control for testing the duplicate samples. I.e. in these cases, the contribution of control variance to the repeatability variance is not considered. Nevertheless, in the absence of alternative options, the reproducibility variance $s^2_R$ of the test systems was calculated according to equation (1), with the repeatability approximated by the duplicate samples of one and the same sediment. These however, were only available for sediment 1, 2 and 3. For sediment 4-7, the total

variance was approximated solely by the inter-laboratory variance $s^2_L$. To keep it simple, it will nevertheless be termed “reproducibility” as described above.

Computation of means and standard deviations according to DIN ISO 5725-2 requires normal distribution and negative values are not acceptable. Therefore, negative percentages of inhibitions (stimulations) were set to zero percent inhibition. The LID data were ln-transformed, since they are derived from dilution series based on factor 2 and the original data therefore show a log normal distribution, rather than a normal distribution. All calculations as described in DIN ISO 5725-2 were performed with the ln-transformed data, subsequently the results were retransformed to the original scale, proceeding as follows:
In a first step, the original LID data were transformed by Y = ln (X), then arithmetic mean μ and the standard deviation σ for Y were calculated[5], with weighting according to the number of repeated measurements as prescribed by DIN ISO 5725-2.

Inverse transformation of μ and σ was performed according to the formula for log normal distribution, resulting in the following parameters for the original scale of X:

Geometric mean [6] = EXP (μ) (2)

95% confidence interval CI[7] = $\mathrm{EXP}\left(\mu \pm \frac{\sigma}{\sqrt{n}} * t_{(0{,}025;\ df)}\right)$ (3)

with t = onesided 2,5% Quantile of t-distribution
and df = n-1

95% and 99% prediction interval PI [8] = EXP ($\mu \pm \sigma * z$) (4)
with z = 1,96 and 2,57 for 95 % and 99% PI, respectively

Expected value EV (X) [9] = $\mathrm{EXP}\left(\mu + \frac{1}{2} * \sigma^2\right)$ (5)

Standard deviation std (X) = $\mathrm{EV}(X) * \sqrt{e^{\sigma^2} - 1}$ (6)

[5] The terms μ and σ are actually used for the theoretical basic population. Here they are used to clearly indicate the arithmetic mean $y\_ = \mu$ and the standard deviation $s_y = \sigma$ of the *log-transformed sample population*, from which all other parameters are derived.
[6] Geometric mean = The mean of a set of *n* values computed by extracting the *n*th root of the product of the values. The logarithm of of the geometric mean equals the arithmetric mean of the logarithms
[7] The 95% confidence interval of the mean is the range, where the calculated mean is to be expected with a probability of 95%.
[8] The 95% (or 99%, respectively) prediction interval of the mean is the range, where a single laboratory result is to be expected with a probability of 95% or 99%, respectively. It is also called tolerance range.
[9] Expected value (EV) = location parameter of a probability distribution (the sum of all events, weighted with the event probability) denotes the expected center of the distribution; the EV is the mean value to which the measured values with increasing sample size and number of experiments tend. In normally distributed variables, the EV equals the arithmetic mean, in lognormal distribution, the EV always exceeds the arithmetic mean.

Coefficient of variation CV% = $\frac{Std\ (X)}{EV\ (X)} * 100$ (7)

All calculations were performed using MS Excel™.
It should be explicitly stated, that retransformation according to equation (6) only gives the value for the overall standard deviation on the original scale, i.e. the reproducibility variance, which is the sum of intra- and interlaboratory variance. It is not possible, to calculate the values for intra- and interlaboratory variance on the original scale by retransformation. Therefore, for variable LID, results for repeatability $s^2r$ and interlaboratory variance $s^2L$ are not available.

### 2.3.2. Normalized Width as reliability indicator

The CV% for reproducibility expresses the variability of a certain variable as percentage of the corresponding mean. This is of low informative value if the variable itself is already expressed as percentage, as it is the case for "%inhibition in D1". Depending on the absolute number of "%inhibition in D1", the corresponding variability expressed in percent becomes very high – especially for low absolute values (e.g. 10 %inhibition in D1 $\pm$ 10%, i.e. a range of 0%-20% inhibition in D1, would correspond to 100% variability).
Therefore, as an additional indicator for reliability, the "normalized width" (NW) was calculated, based on the mean and its 95% confidence interval. The 95% CI gives the range, where the "true mean" lies with a probability of 95%. Adapting the concept proposed by the EFSA Technical Report 2015 [7]) for ECx values, the NW is calculated as the ratio between the width of the 95% confidence interval and the mean:

NW = $\frac{95\%CI\ upp - 95\%CI\ low}{mean}$ (8)

The higher the NW, the higher is the uncertainty of the corresponding mean. E.g. a value of NW = 2 means that the true mean will be within a range of twice the calculated mean E.g. if the mean %inhibition in D1 is 40% with NW = 2, the 95%CI is 2x40%" = 80%, i.e. 40% $\pm$ 40 percent points, i.e. 0%-80%. NW values below 1 indicate a low variability: e.g. if the mean %inhibition in D1 is 40% with NW = 0.5, the 95%CI of the mean "%inhibition in D1" is 0.5x40%" = 20%, i.e. 40% $\pm$ 10% percent points, i.e. 30%-50%.
Thereby, in case of the variable "%inhibition in D1", the 95% confidence interval of the mean is symmetric, i.e. the upper and lower range can be directly calculated based on the mean and its NW. That does not hold for the variable LID because of the asymmetric nature of the 95% CI for lognormally distributed variables. Hence, if the mean LID e.g. is 4 and the NW is 2, the true mean LID will be within a range of 4x2=8 LID categories. But this does not necessarily mean a range from 0-8 but, may also be e.g. 2-10.

### 2.3.3. Outlier analysis

According to DIN ISO 5725-2, outlier analysis should be performed both for single measurements per laboratory and for laboratory means. However, in view of the limited data base, the present evaluation refrained from outlier analysis by means of certain statistical tests. Nevertheless, it was checked by the so called "warning charts", whether the laboratory specific geometric means (or single results, if no repeated measurements were performed) are supposed to be outliers: The 95%- and 99%-prediction intervals (also called tolerance ranges) of the overall mean represent the range, in which the result of a single laboratory will fall with 95% and 99% probability (see Eq 4 presented above). Outlying data sets can be easily identified by plotting the overall mean and these two tolerance limits with the single laboratory results[10] [5]. If a single laboratory result falls beyond the 99%-PI, it is defined as a "statistical outlier". A result located between the two PIs is termed "struggler" (nearly outlier). For the Algae Growth Inhibition test, 4 laboratory specific results were found to be "nearly outliers", for the Bacteria Luminescence test, this applied to 3 values[11]. No values were excluded for further evaluation.

### 2.3.4. Individual Laboratory Performance – scores

The objective "correctness of the results" of the AGI or the LBT cannot be determined since the real toxicity of the tested sediments is not known. Instead, a reference value can be approximated by the overall mean of the results of all measurements of a certain sediment. To assess the performance of an individual laboratory by means of a score, its individual result for a certain sample is related to the reference value.

To standardize the deviation of a single result from the overall mean of a sample, the calculation of z-scores is described in DIN 38402-45 [4]: For each sample, the overall mean and the reproducibility standard deviation are determined and used as reference values to calculate the z-score as follows:

$$\text{z-Score} = (\text{result}_{ind} - m_{ref})/s_{ref} \quad (9)$$

with

$\text{result}_{ind}$ = individual result; here: LID or %inhibition in D1

reference mean $m_{ref}$ = mean of all results; here: mean LID or %inhibition in D1

reference standard deviation $s_{ref}$ = reproducibility $s_R$, here: sR of LID or % inhibition in D1

[10] "Warning chart concept", Guidance Document on Statistical Methods for Environmental Toxicity Tests. Report EPS 1/RM/46, Environment Canada, 2005

[11] Identified as nearly outliers, i.e. beyond the 95% PI:

AGI: (1) EL, all data, LID, Sed 3, L04 (LID=32); (2) PW, all data, %Inhib, Sed2, L05 (29,5%, invalid); (3) PW, only valid data, %Inhib, Sed2, L06 (16,8%); (4) PW, only valid data, %Inhib, Sed3, L05 (32,5%);

LBT: (1) EL, all data, %Inhib, Sed 3, L03 (15,7%); (2) PW, all data, LID, Sed2, L03 (LID=4); (3) PW, only valid data, LID, Sed2, L03 (LID=4)

According to the definition of the prediction interval (see above, Eq. 4), results with a z-score higher than $\pm$ 2, can be assessed as nearly outlier with an error probability of 5%, since assuming normal distribution of the test results, the probability that z is lower or equal 2, is about 95%.

According to DIN 38402-45, the calculations should account for the effect of possible outliers. Therefore, the overall mean used as reference value and the reproducibility used as reference standard deviation should be determined by robust methods, such as Hampel estimator or Q-method. Moreover, DIN 38402-45 states, that the informative value of the resulting z-scores becomes highly questionable, if the estimation of the reference standard deviation is based on only few laboratories and if several laboratories show identical results.
So, in view of the very small data base and the above mentioned limitations, the present evaluation refrains from calculation of z-scores strictly according to DIN38402-45. Instead, the principle way of calculating a score as a standardized measure for the performance of the individual laboratory results is applied in a simplified way: the reference value and reference standard deviation given in Eq. (9) are approximated by the overall means and reproducibilities obtained in the first part of the present evaluation (see section 2.3.1.). To prevent miss-interpretations, the resulting scores are called pseudo-z-scores and any results and conclusions should be regarded only as tendencies.
The pseudo-z-scores are used to assess whether a laboratory tends to *systematically* overestimate or underestimate toxicity compared to other laboratories: A score = 0 means that the individual laboratory result is identical with the overall mean. Assuming a random variation of the results, the scores of a single laboratory obtained for different samples should equally show positive and negative deviations from zero, and thus the *mean score* of a single laboratory should equal zero. The degree of deviation of the mean score of a laboratory from zero indicates the degree of systematic overestimation or underestimation of toxicity by this laboratory.

Since LID data are log normal distributed, all calculations were performed with ln-transformed data. For %inhibition in D1, the original percentages were used, with negative inhibitions, i.e. stimulations, replaced by zero.

# 3. Results

In section 3.1., the results for toxicity parameters LID and %inhibition in D1 (i.e. in the undiluted sample), obtained in AGI and LBT using both eluate and pore water samples, are presented. If eluate and pore water samples provide different LIDs, the higher LID is used to assess the overall toxicity of the corresponding sediment. The same applies to the mean LIDs, which are calculated based on only results from eluate and only results from pore water samples.
In section 3.2, relations between the LID obtained in a certain test and the %inhibition measured in D1 will be presented. Section 3.3. will show the individual laboratory performance for each biotest by means of percentages of valid tests performed and by pseudo-z-scores.

Generally, both results based on all data and on only valid data are presented. Results based on only valid data were sometimes slightly different from those based on all data, but basically in the same range. Therefore, the following presentations focus on results from valid data.

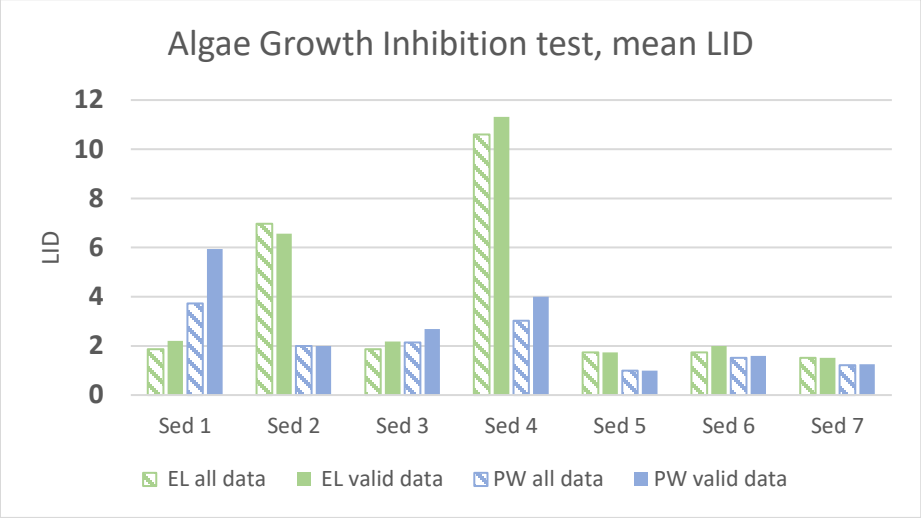
## 3.1. Toxicity Parameter

### 3.1.1. Algae Growth Inhibition test

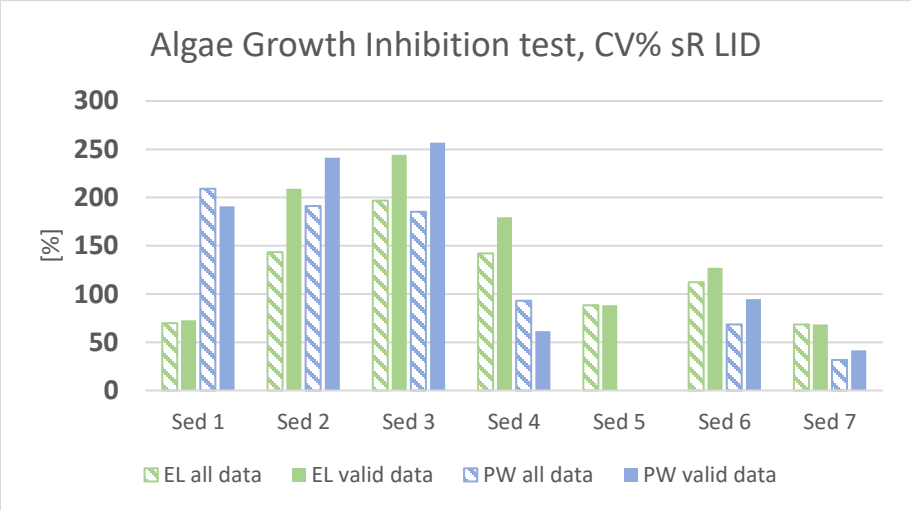
Figure 1 and Figure 2 show the means and CV% for reproducibilities obtained for the variables "LID" (Figure 1) and "Percent Inhibition in D1" (Figure 2) in AGI for eluate (green bars) and pore water samples (blue bars). The detailed results are available in Annex Tables 3 - 6. Results based on all data (hatched bars) and on only valid data (solid bars) were quite similar.
Sediment 4 was found to show the highest mean **LID** and thus toxicity (mean LID around 10), sediment 1 and sediment 2 showed mean LIDs around 5 and 7, respectively and thus moderate toxicity, and low mean LIDs and thus low toxicity was found for sediments 3, 5, 6 and 7 (mean LIDs 1-2). Nevertheless, the broad 95% confidence ranges of the means indicate their high uncertainty (e.g. sed 1: mean LID = 5.9 (95% CI 0.8 - 42.7), sed 2: mean LID = 6.5, 95% CI = 0.8 – 51.7), sed 4: mean LID = 11.3 (95% CI 1.7 - 76.4), sed 3, 5, 6 and 7: mean LID = 1.0 – 2.7 (95% CIs between 0.2 - 26 (AGI). This is both due to high standard deviation and low sample size, see equation (3).
When comparing results from **eluate and pore water samples**, three sediments (sed 1, 2 and 4), revealed markedly different mean LIDs and thus toxicity in eluate and pore water samples, with two times eluate showing the higher toxicity (sed 2 and 4) and one time pore water showing the higher toxicity (sed 1). In four sediments (sed 3,5,6 and 7), the LIDs measured in eluate and pore water samples were quite similar in each case.
**Repeatability** can be assessed by the duplicate samples. In 26 out of 30 duplicate samples of sediment 1, 2 and 3, identical LIDs (20) or LIDs which differ not more than one dilution step (6) were obtained (Table 2, Annex Table 3 and 4).
**Reproducibilites** of the LIDs in most cases were about 100% or even higher (Fig. 2, Annex Table 3 and 4). The obtained mean LIDs frequently show NW-values > 2, and sometimes even up to 10 (Annex Table 3 and 4).

Sediment 4 showed the highest mean **%inhibition in D1** (about 70%). In sediment 1 and 2 the mean %inhibitions in D1 were up to 45% and in sediment 3, 5, 6 and 7 the mean %inhibition in D1 was below 30%. Thereby, the results of **eluate and pore water** agreed quite well in sediment 4, 5, 6 and 7, whereas in sediment 1, 2 and 3, generally the eluate samples showed markedly higher mean %inhibition in D1.
**Repeatability** CV% for variable %inhibition in D1 showed no clear tendency: whereas for eluate samples, CV% sr was around 20%, for pore water samples CV% sr was between 0% and 280%.
The CV% for **reproducibility** frequently was around or higher 100%. The values for NW were frequently around 2 and in some cases up to 5.5 at maximum, indicating a high, but somewhat lower uncertainty than found for the variable LID.

**Figure 1**: Mean and CV% for reproducibility $s_R$ of LID obtained in the Algae Growth Inhibition test; EL = Eluate (green), PW = Pore Water (blue); sample sizes: see Annex Table 3 and 4. Hatched bars: all data; solid bars: only valid data.

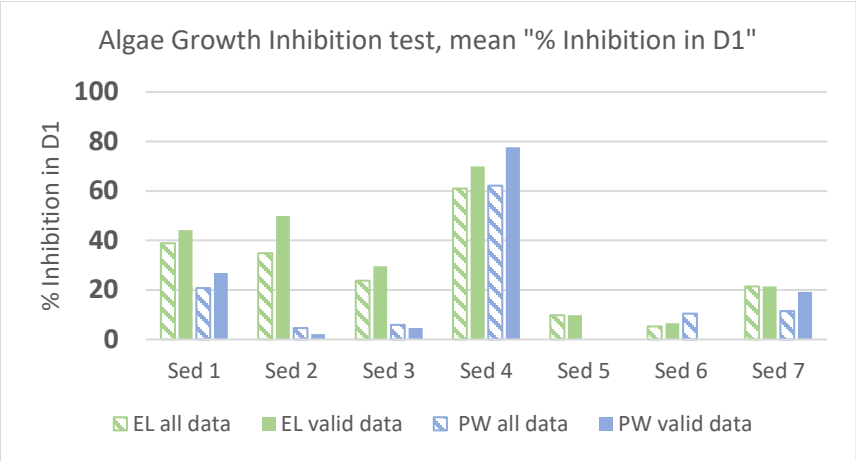



| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 1,87 | 6,97 | 1,87 | 10,60 | 1,74 | 1,74 | 1,52 |
| EL valid data | 2,21 | 6,56 | 2,2 | 11,31 | 1,74 | 2,00 | 1,52 |
| PW all data | 3,73 | 2,00 | 2,14 | 3,03 | 1,00 | 1,52 | 1,21 |
| PW valid data | 5,94 | 2,00 | 2,7 | 4,00 | 1,00 | 1,59 | 1,26 |

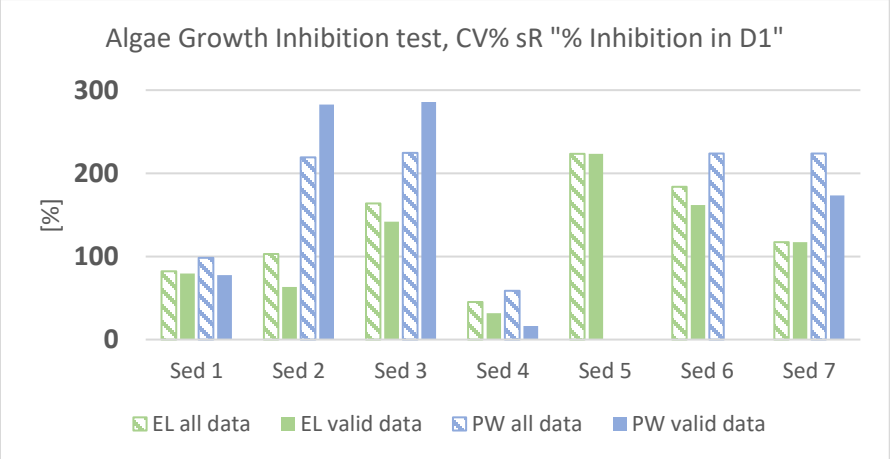



| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 69,70 | 143,3 | 196,90 | 142,10 | 88,20 | 112,30 | 68,40 |
| EL valid data | 72,9 | 209,1 | 244,2 | 179,6 | 88,2 | 127,0 | 68,4 |
| PW all data | 209,1 | 191,0 | 185,20 | 93,10 | 0,00 | 68,40 | 31,70 |
| PW valid data | 190,8 | 241,3 | 256,7 | 61,4 | 0,0 | 94,7 | 41,6 |

**Figure 2**: Mean and CV% for reproducibility $s_R$ of "Percent Inhbition in D12 obtained in the Algae Growth Inhibition test, EL = Eluate (green), PW = Pore Water (blue); sample sizes: see Annex Table 5 and 6. Hatched bars: all data; solid bars: only valid data. Please note, that the CV% for variable "Percent inhibition in D1" is an artefact due to specific feature of the variable, see text.




| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 38,90 | 34,9 | 23,70 | 61,00 | 9,80 | 5,20 | 21,40 |
| EL valid data | 44,20 | 49,9 | 29,6 | 69,90 | 9,80 | 6,60 | 21,40 |
| PW all data | 20,8 | 4,6 | 5,9 | 62,1 | 0,0 | 10,4 | 11,5 |
| PW valid data | 26,8 | 2,1 | 4,6 | 77,7 | 0,0 | 0,0 | 19,2 |



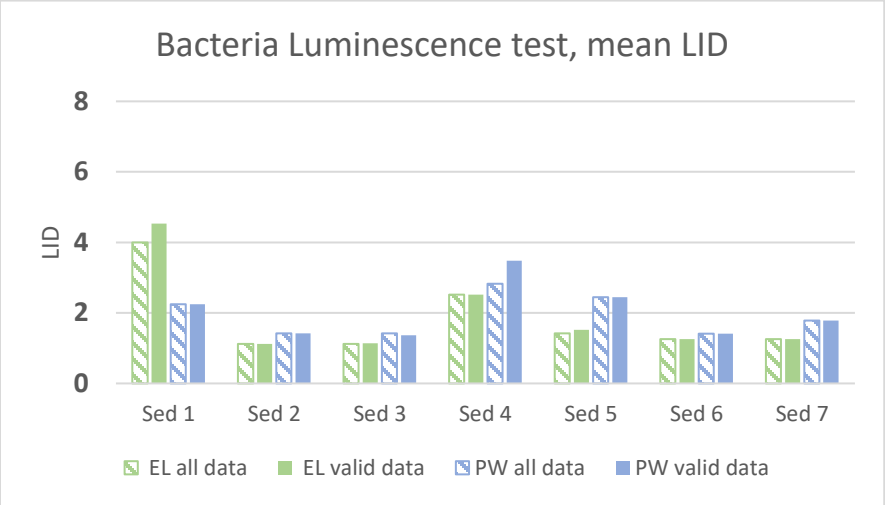

| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 82,20 | 102,9 | 163,60 | 45,30 | 223,50 | 183,90 | 117,20 |
| EL valid data | 79,6 | 63,4 | 142,0 | 31,8 | 223,5 | 161,9 | 117,2 |
| PW all data | 98,2 | 219,0 | 224,5 | 58,6 | 0,0 | 223,6 | 223,6 |
| PW valid data | 77,7 | 282,8 | 285,8 | 16,3 | 0,0 | 0,0 | 173,2 |

### 3.1.2. Bacteria Luminescence test

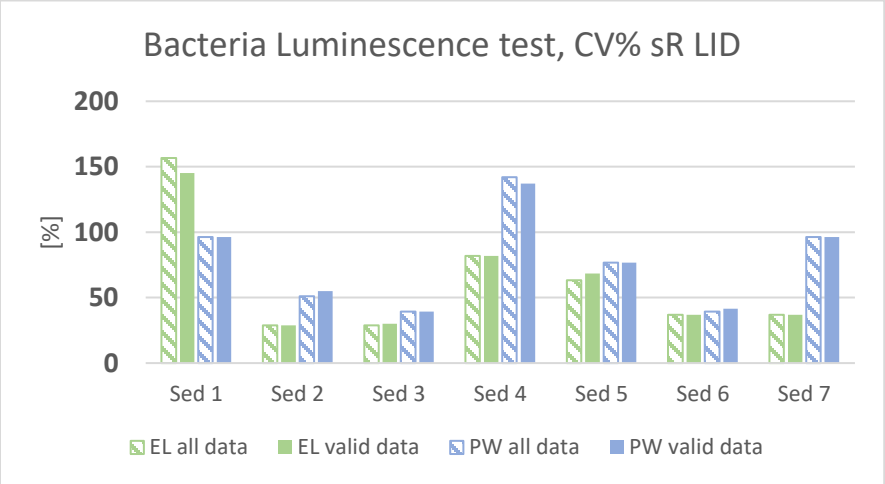
Figure 3 and Figure 4 show the means and reproducibilities obtained for the variables “LID” (Figure 3) and “Percent Inhibition in D1” (Figure 4) in the LBT for eluate (green bars) and pore water samples (blue bars). The detailed results are available in Annex tables 7 - 10. Results based on all data (hatched bars) were always in high agreement with those based on only valid data (solid bars).
Sediment 1 and 4 were found to be of moderate toxicity (mean **LIDs** about 3-4), all other sediments were found to be of low toxicity or even non toxic (mean LIDs 1-2). Thereby, the higher the mean LID, the higher is its uncertainty (sed 1: mean LID = 4.5 (95%CI = 1.5 - 13.9); sed 4: mean LID =3.5 (LBT, 95%CI 1.0 - 12.5). Sed 3, 5, 6 and 7 mean LID = 1.1 – 2.5, with the corresponding 95% CIs between 0.7 - 4.6 ), sed 2: mean LID = 1.4 , (95% CI 0.7 – 2.7).
Generally, **eluate and pore water samples** revealed very similar mean LIDs, except for sed 1 where the mean LID found in eluate was about twice that of pore water.
**Repeatability** of LID can be assessed by the duplicate samples. In all 36 duplicate samples of sediment 1, 2 and 3, identical LIDs (32) or LIDs which differ not more than one dilution step (4) were obtained (Table 3, Annex Table 7 and 8).
The **reproducibilities** show a different picture: in four cases, the reproducibilities of LID were around 100% or exceeded 100%, but also in six cases reproducibilities around or lower 40% were obtained. In four cases, the reproducibilities CV%sR were between 50% and 80%. Thereby, the lower the toxicity (i.e. the mean LID), the better its reproducibility: in the sediments with low toxicity, the obtained mean LIDs frequently show NW-values below 1 or up to 2 at maximum. In the sediments with higher toxicity, NW values up to 11.5 were found.

Sediment 1 and 4 showed the highest mean **%inhibition in D1** (about 30%-40%), in sediment 5, 6 and 7 the mean %inhibitions in D1 were 20% $\pm$ 10% and in sediment 2 and 3 the mean %inhibition in D1 was below 20%. Thereby, **eluate and pore water samples** showed quite similar results with a tendency to slightly higher %inhibitions in D1 in pore water.
**Repeatability** CV% for variable %inhibition in D1 showed no clear tendency: whereas for eluate samples, CV% sr was between 35% and 57%%, for pore water samples CV% sr was much lower (between 6% and 14%).
The **reproducibility** of %inhibition in D1 in most cases was around 80% or even higher. The values for NW were frequently around 2 and in some cases up to 3.3 at maximum.

**Figure 3**: Mean and CV% for reproducibility of LID obtained in the Bacteria Luminescence test. EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data. Sample sizes: see Annex Table 7 and 8.

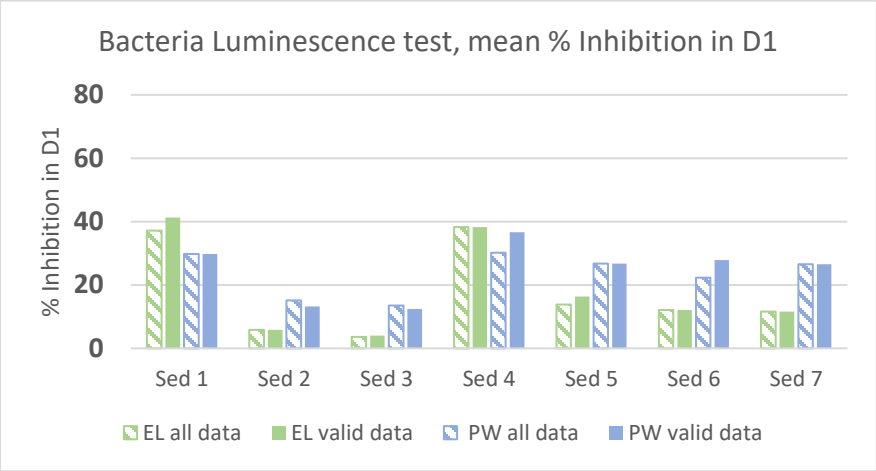



| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 4,00 | 1,12 | 1,12 | 2,52 | 1,42 | 1,26 | 1,26 |
| EL valid data | 4,53 | 1,12 | 1,1 | 2,52 | 1,52 | 1,26 | 1,26 |
| PW all data | 2,25 | 1,42 | 1,42 | 2,83 | 2,45 | 1,41 | 1,78 |
| PW valid data | 2,25 | 1,42 | 1,37 | 3,48 | 2,45 | 1,41 | 1,78 |

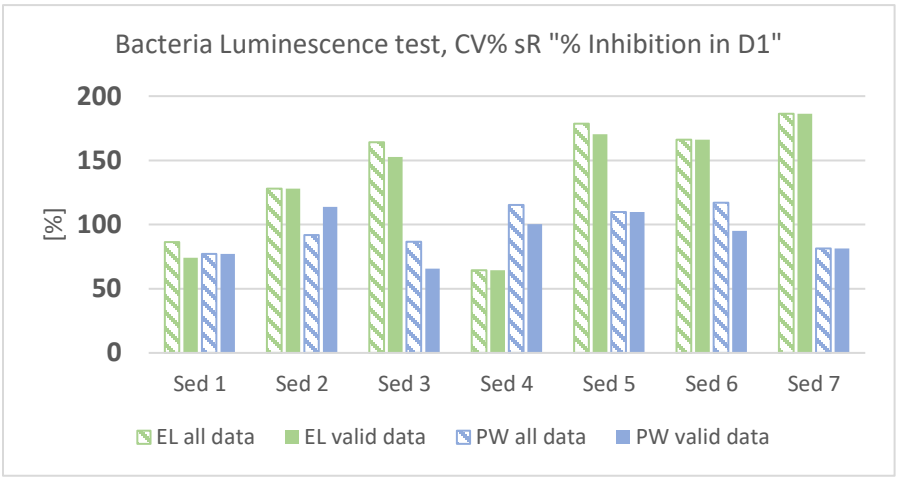



| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 156,7 | 28,8 | 28,8 | 81,8 | 63,2 | 36,9 | 36,9 |
| EL valid data | 145,1 | 28,8 | 30,0 | 81,8 | 68,4 | 36,9 | 36,9 |
| PW all data | 96,3 | 51,1 | 39,3 | 142,0 | 76,8 | 39,3 | 96,3 |
| PW valid data | 96,3 | 54,9 | 39,3 | 137,0 | 76,8 | 41,6 | 96,3 |

**Figure 4:** Mean and CV% of reproducibility of percent inhibition in D1 obtained in the Bacteria Luminescence test. EL = Eluate (green), PW = Pore Water (blue); Hatched bars: all data; solid bars: only valid data. Sample sizes: see Annex Table 9 and 10. Please note, that the CV% for variable "Percent inhibition in D1" is an artefact due to specific feature of the variable, see text.




| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 37,2 | 5,8 | 3,6 | 38,3 | 13,8 | 12,1 | 11,6 |
| EL valid data | 41,3 | 5,8 | 4,0 | 38,3 | 16,3 | 12,1 | 11,6 |
| PW all data | 29,8 | 15,1 | 13,5 | 30,2 | 26,8 | 22,3 | 26,5 |
| PW valid data | 29,8 | 13,2 | 12,4 | 36,6 | 26,8 | 27,9 | 26,5 |




| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
|---|---|---|---|---|---|---|---|
| EL all data | 86,3 | 128,0 | 164,2 | 64,3 | 178,6 | 166,0 | 186,4 |
| EL valid data | 74,2 | 128,0 | 152,6 | 64,3 | 170,3 | 166,0 | 186,4 |
| PW all data | 77,1 | 91,8 | 86,5 | 115,3 | 109,8 | 117,1 | 81,4 |
| PW valid data | 77,1 | 113,7 | 65,6 | 100,2 | 109,8 | 95,0 | 81,4 |

### 3.1.3. Comparison AGI – LBT

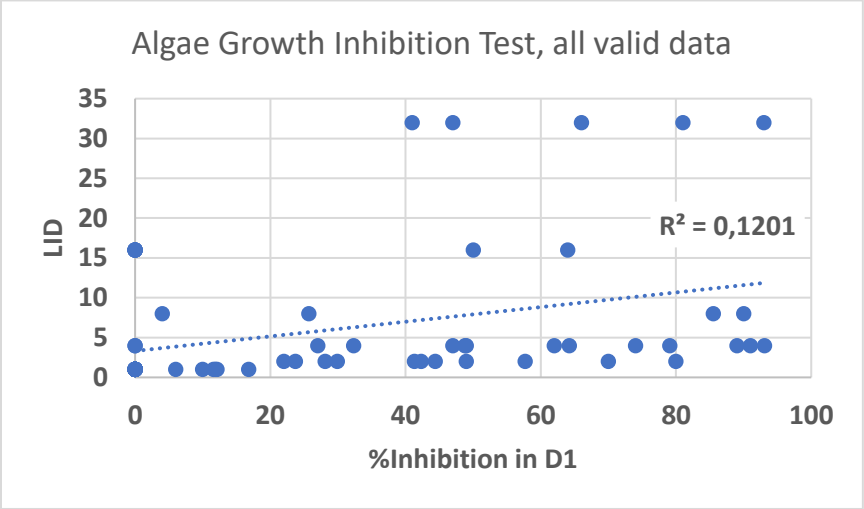
Generally, the overall levels of LIDs detected by the LBT were lower compared to the results of the AGI or equal in case of anyway low toxicity detected by the AGI. Hence, the LBT proved to be less sensitive to toxicity, detects another mode of action to toxicity or is less susceptible to cofounding factors (e.g. ammonium or nutrients), than the AGI.
Based on mean LIDs, Sediment 1 and 4 were identified as the most toxic sediments in the round robin test both by AGI and LBT (sediment 1: mean LID = 5,9 (AGI) and 4,5 (LBT); sed 4: mean LID = 11,3 (AGI) and 3,5 (LBT); results from valid data). Sediment 3, 5, 6 and 7 where consistently found to be of low toxicity or non toxic in both tests (mean LID = 1,0-2,7 (AGI) and mean LID = 1,1 – 2,5 (LBT); results from valid data). For sediment 2, there were no consistent results: while the mean LID from LBT was 1,4, indicating very low toxicity, the mean LID from AGI indicates moderate toxicity (LID = 6,5) (results from valid data). However, the reproducibility of the obtained mean LIDs generally is very low, i.e. the reliability of the obtained mean LIDs is poor.
The LBT basically showed a better reproducibility for variable LID than the AGI, but nevertheless a broad range of variability between 30% and higher than 100%.
Repeatability for variable LID was assessed by the duplicate samples. Both in AGI and in LBT, 26 out of 30 (AGI) or all 36 (LBT) duplicate samples, respectively, showed identical LIDs or LIDs differing not more than 1 dilution step.
In contrast to the LID, the mean %inhibition in D1 found in AGI and LBT in the tested sediments showed no consistent pattern: in some tests, the %inhibition in D1 obtained in the AGI exceeded that of the LBT, sometimes it was the other way round, and in some cases they were similar.
The reproducibility of %inhibition in D1 in most cases was around 80% or even higher, with the AGI showing the higher percentages. If reproducibility of %inhibition in D1 is assessed by normalized width (NW), the values for NW are frequently around 2 and in some cases up to 5.5 at maximum (AGI) or 3.3 at maximum (LBT), indicating a somewhat lower uncertainty for the variable %inhibition in D1 than found for the variable LID.
Repeatability CV% sr for %inhibition in D1 in most cases of AGI and LBT was lower than 50% or even lower than 20%.

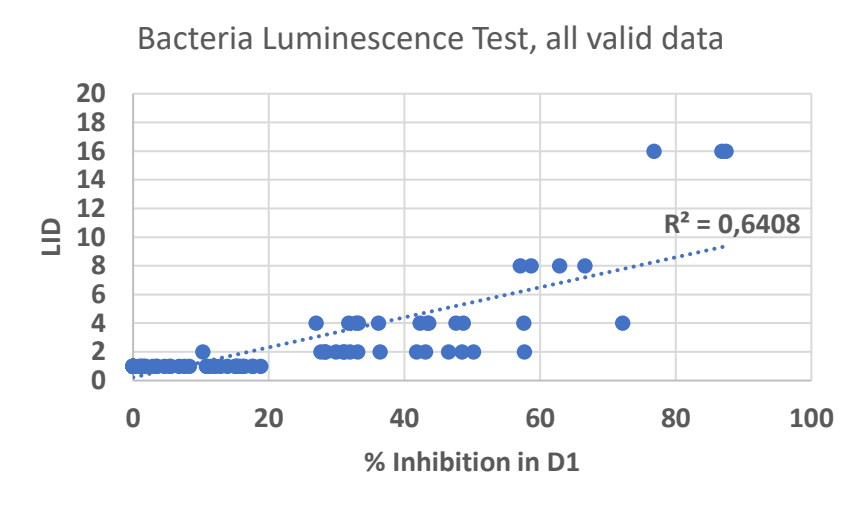
## 3.2. Relationship between LID and % inhibition in D1

Regardless of the observed uncertainties, the obtained *mean* LIDs for a certain sample and the corresponding *mean* %inhibition in D1 reveal quite consistent results both in AGI and in LBT. When analyzing the *individual results per laboratory*, however, a different picture is obtained. Figure 5 shows that for Bacteria Luminescence test, there is a correlation between both variables ($R^2$ = 0,64), i.e. the higher the %inhibition measured in the undiluted sample, the higher is the LID obtained from the complete dilution series for the corresponding sediment (within a precision of two dilutions). In contrast, no clear relationship between “%inhibition in D1” and LID was found for the Algae Growth Inhibition test ($R^2$ = 0,12): regardless of the % inhibition measured in the undiluted sample, LIDs

between 1 and 32 were obtained for the corresponding sediment.
This might be related to the phenomenon of stimulating effects observed in the tested sediments. In 53% of Algae Growth Inhibition tests (53 out of 100) and in 29% of the Bacteria Luminescence tests (20 out of 100) a stimulation was found in D1 (see Annex tables 1 and 2). Stimulation might be caused by enhanced nutrient concentrations. The higher proportion of stimulations in the AGI might be due to the longer test duration and thus exposure time of the AGI compared to the LBT.
Stimulating effects probably compensate for potential toxic effects. Since they are not restricted to the undiluted sample (D1), stimulating effects may cause higher variability and thus hamper the determination of LID as toxic parameter. This might explain the lacking correlation between %inhibition in D1 and LID for the Algae Growth Inhibition test.

**Figure 5:** Correlation between laboratory specific values for LID and %inhibition in D1 obtained in Algae Growth Inhibition Test and Bacteria Luminescence Test. Data base: all valid single test results both from eluate and pore water samples. Negative inhibitions (i.e. stimulations) were set to zero (see Table 2 and Table 3). Please note, that various data points with identical numbers (e.g. LID = 16, %inhibition = 0), appear as only one dot in the figures.



## 3.3. Individual Laboratory Performance

The individual laboratory performance can be assessed by the ability to perform a valid test and by the standardized deviation of the single laboratory results from the corresponding overall mean result, assessed by pseudo-z-sores.

### 3.3.1. Ability to perform a valid test

Since the same controls were used for multiple tests by several labs, this analysis is based on the percentage of valid controls, rather than on valid tests. The percentage of valid controls per laboratory and per biotest is presented in Table 6. For the Algae Growth Inhibition test, 14 out 36 controls failed to meet the validity criteria, corresponding to 39%. Thereby, all invalid controls were derived by L05 (4 out of 10 controls were invalid, i.e. 40% of the controls run by L05) and L06 (10 out of 14 controls were invalid, i.e. 71% of the controls run by L06).
For the Bacteria Luminescence test, 9 out of 62 controls were found to be invalid, corresponding to 15%. L06 accounted for most of the invalid controls (7 out of 20, i.e. 35% of the controls run by L06).

**Table 6:** Percentage of invalid controls per laboratory in all Algae Growth Inhibition tests (AGI) and all Luminescence Bacteria tests (LBT). Please note, that the invalid LBT of L04 is due to high variability in a treatment, rather than in the control. It is nevertheless considered in the percentage of invalid controls, as the latter here is just a measure for the ability of the laboratories, to perform a valid test.

| AGI, EL and PW | | | |
|---|---|---|---|
| | controls | invalid | % |
| L1 | 6 | 0 | 0% |
| L2 | 2 | 0 | 0% |
| L3 | - | - | - |
| L4 | 4 | 0 | 0% |
| L5 | 10 | 4 | 40% |
| L6 | 14 | 10 | 71% |
| **total** | **36** | **14** | **39%** |

| BLT, EL and PW | | | |
|---|---|---|---|
| | controls | invalid | % |
| L1 | 7 | 0 | 0% |
| L2 | 6 | 0 | 0% |
| L3 | 4 | 0 | 0% |
| L4 | 5 | 1* | 20% |
| L5 | 20 | 1 | 5% |
| L6 | 20 | 7 | 35% |
| **total** | **62** | **9** | **15%** |

* test invalid because of too high variability in a treatment

### 3.3.2. Pseudo z-scores

For every laboratory, the pseudo-z-scores for the variables LID and %inhibition in D1 obtained in the tested sediments were calculated. The results are shown in Figure 6 (AGI) and Figure 7 (LBT). Only valid data were considered, since the prerequisite for further assessing the quality of a result is its validity.

**1. Absolute values of the individual scores.**
Individual scores between -1 and +1 indicate that the corresponding result is within 68% tolerance range of the reference value and thus unremarkable. Only individual scores close to or beyond -2 or +2 (i.e. the

95% tolerance range of the reference value) indicate possible issues with laboratory performance.
In the present round robin test, most scores were unremarkable, with few exceptions found for L03, L04 und L05 (Figure 6 and Figure 7). However, as pointed out before, due to the limited data base, the pseudo-z-scores presented in this report should be just regarded as tendencies to get a rough impression of the individual laboratory performance.

**2. Direction of the individual scores.**
Assuming a random variation of the results, the scores of a single laboratory for a certain variable obtained for different samples should roughly show an equal number of positive and negative deviations from zero. So, regardless of the absolute values of the individual scores of a single laboratory, their *mean* should be close to zero. The higher the deviation of the mean score of a laboratory from zero, the higher the probability that there is a *systematical* overestimation or underestimation of toxicity by this laboratory.
From a visual inspection, Figure 6 and Figure 7 reveal, that for the Algae Growth Inhibition test, the scores show no obvious pattern of systematic tendencies. For the Bacteria Luminescence test, the scores of L03, L04 and L05 indicate that there may be a systematic pattern.

This can be verified by the mean scores for every laboratory presented in Figure 8, per variable and biotest. Mean scores in most cases were between -0.5 and + 0.5, indicating a more or less random scattering of the laboratory specific results. In contrast, for L03 there was evidence that the LBT tends to detect increased values for toxicity measures compared to other laboratories (mean score LID = 0.972, mean score %inhibition in D1 = 1.371). The same applies to L05 (LBT, mean score LID 0.872)[12], whereas for L04, the mean scores point to reduced values for toxicity in the LBT (mean score LID = - 0.691; mean score %inhibition in D1 = -1.441) and slightly increased values of toxicity in the AGI (mean score LID = 0.606).

[12] The elevated pseudo-z-score for L05 in the LBT may be at least partly due to the fact, that the lab didn´t test D1 and thus LIDs could be only given as $\leq$ 2. These were set to 2 for statistical evaluation, thus, "2" was the lowest LID considered for L05 in the LBT. The resulting pseudo-z-scores thus might be an artefact due to data modification.

**Figure 6:** Laboratory specific pseudo-z-scores obtained for the various sediments in the Algae Growth Inhibition test for variable LID and %inhibition in D1; PW = Pore water, EL = Eluate; data base: only valid data. Orange bars: duplicate sample of a certain sediment.

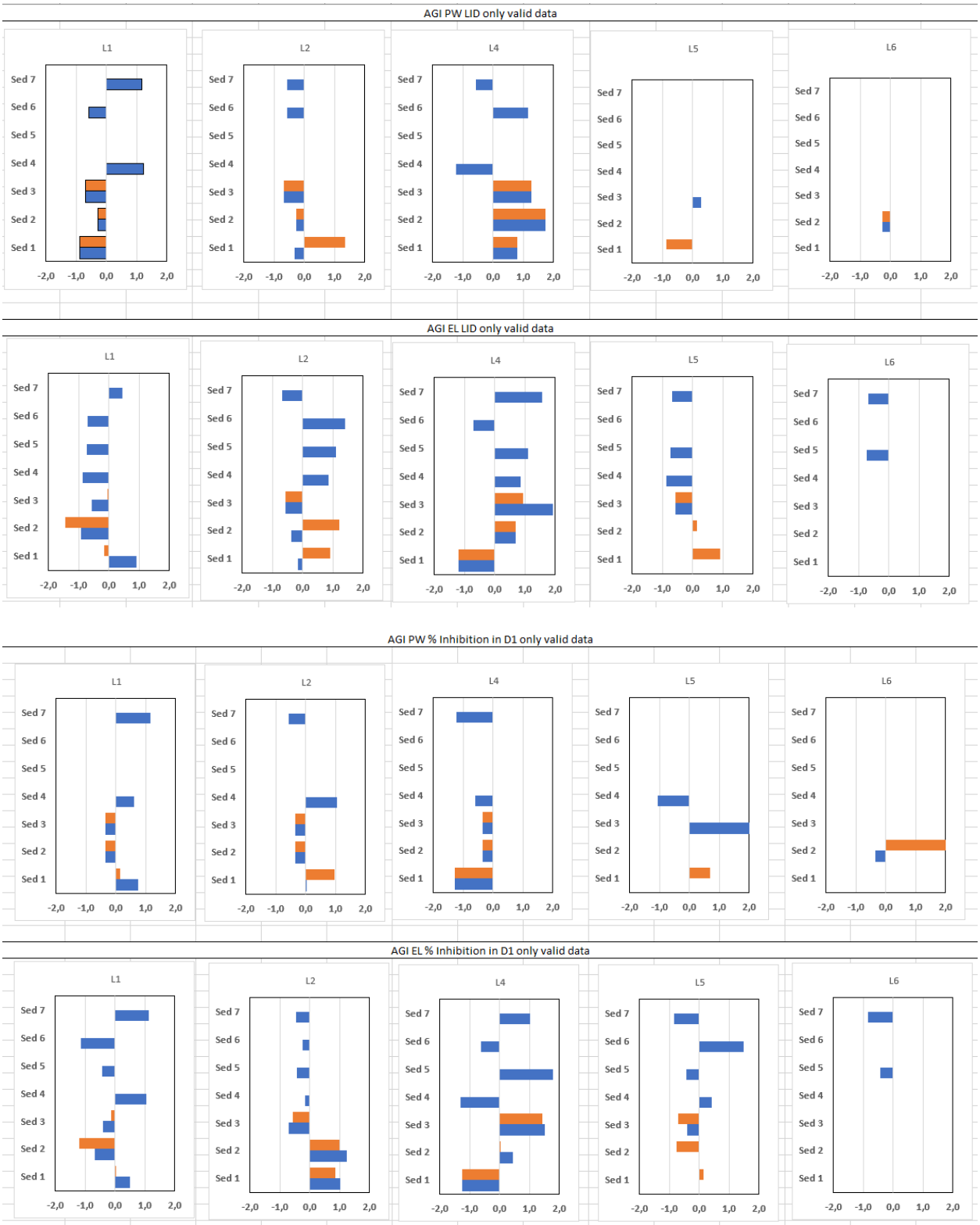


**Figure 7:** Laboratory specific pseudo-z-scores obtained for the various sediments in the Bacteria Luminescence test for variable LID and %inhibition in D1; PW = Pore water, EL = Eluate; data base: only valid data. Orange bars: duplicate sample of a certain sediment.

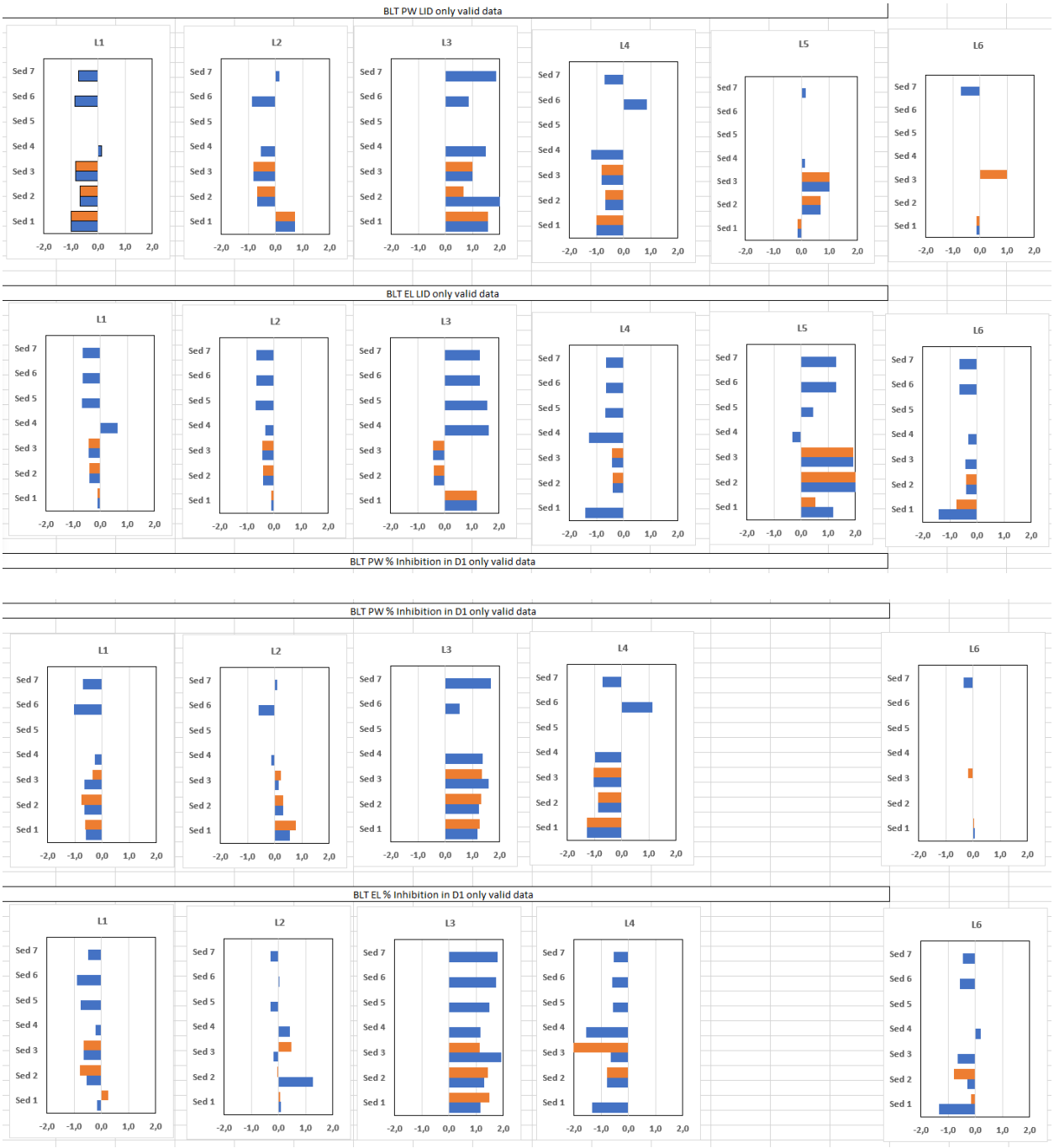
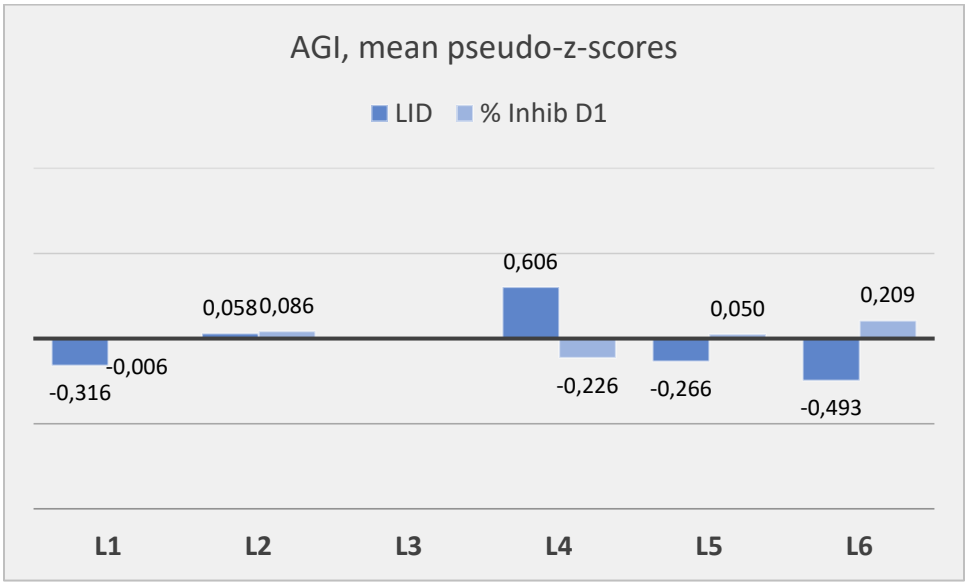
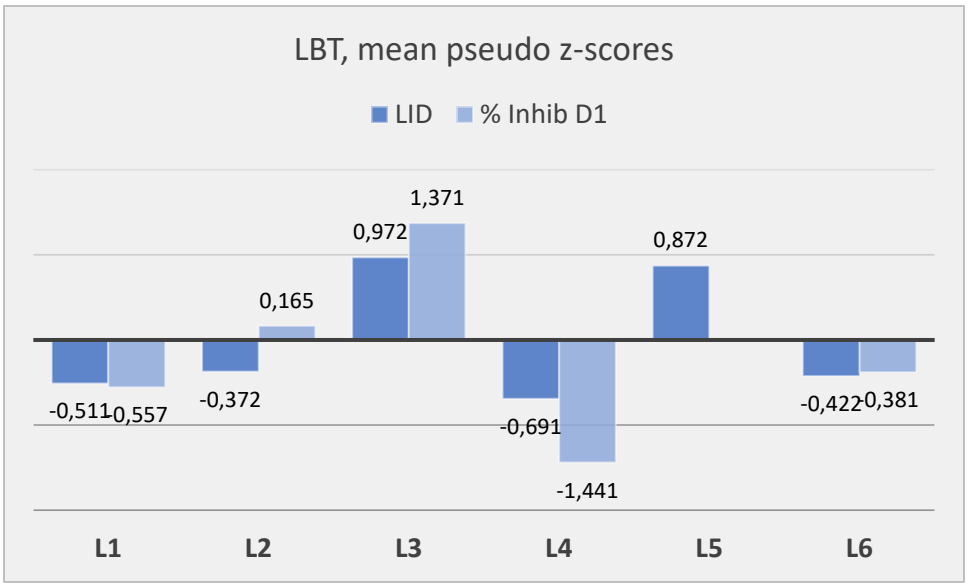


**Figure 8:** Mean laboratory specific pseudo-z-scores obtained for variable LID (dark blue) and %inhibition in D1 (light blue) in the Algae Growth Inhibition test (AGI) and the Bacteria Luminescence test (LBT). Data base: only valid data. Sample sizes indicated in the tables.




| | L1 | L2 | L3 | L4 | L5 | L6 |
|---|---|---|---|---|---|---|
| n LID | 19 | 19 | | 19 | 11 | 4 |
| n %Inh | 18 | 18 | | 18 | 11 | 4 |




| | L1 | L2 | L3 | L4 | L5 | L6 |
|---|---|---|---|---|---|---|
| n LID | 19 | 19 | 19 | 18 | 18 | 12 |
| n %Inh | 19 | 19 | 19 | 18 | | 12 |

# 4. Conclusions

In the following, the initially named objectives of the round robin test and the statistical evaluation should be addressed. In view of the small data base and the limitations of the available data, the following conclusions should be just regarded as tendencies. Results based on all data were sometimes slightly different from those based on only valid data, but basically in the same range. The following conclusions therefore are based on valid data.

**Variability of the measurement variables.**
Repeatability of LID was assessed by the duplicate samples. In the AGI, in 26 out of 30 duplicate samples identical LIDs (20) or LIDs which differ not more than one dilution step (6) were obtained. In the LBT, in all 36 duplicate samples identical LIDs (32) or LIDs which differ not more than one dilution step (4) were obtained.
Reproducibility CV% sR of LID in the AGI in most cases exceeded 100% and NWs were up to 10. In the LBT, some mean LIDs also showed CV% sR up to 100% and NWs up to 10, but generally the LBT provided better reproducibility than the AGI.
Repeatability CV% sr for "%inhibition in D1" in most cases of AGI and LBT was lower than 50% or even lower than 20%.
The reproducibility CV% sR of "%inhibition in D1" in most cases were around 80% or even higher, with the AGI showing the higher percentages. In most cases, values for NW between 2 and 3 were obtained, indicating a generally better reliability of "%inhibition in D1" than of LID.

**Are the AGI and LBT suitable to assess toxicity of different sediments?**
In terms of mean LIDs, both AGI and LBT showed highly variable, but nevertheless consistent results regarding the general toxicity level measured, except for one sediment (sed 2). Thereby, the overall level of mean LIDs detected by the LBT was basically lower compared to those of the AGI or equal in case of anyway low toxicity detected by the AGI. Generally, the precision of the obtained mean LIDs is very low and thus the reliability of of a single result from a single laboratory also is found to be very poor. Thereby, the results from LBT were less variable than those from AGI. This might be due to the overall lower sensitivity of the LBT and / or to lower degree of interferences due to cofounding factors.

**Relationship between the LID and the percentage of inhibition in D1**
Regarding the overall means per sediment, the obtained mean LID for a certain sample and the corresponding "%inhibition in D1" revealed quite consistent results both in AGI and in LBT, i.e. the higher the "%inhibition in D1", the higher the LID.
When analyzing the individual results per laboratory, however, a different picture was obtained. Whereas for the Bacteria Luminescence test, there was a clear correlation between both variables ($R^2$ = 0.64), no clear relationship between "%inhibition in D1" and LID was found for the Algae Growth Inhibition test ($R^2$ = 0.12).

**Performance of the participating laboratories.**
The individual laboratory performance was assessed by the ability to perform a valid test and by the

standardized deviation of the single laboratory results from the corresponding overall mean result, assessed by pseudo-z-sores.

For the Algae Growth Inhibition test, 14 out 36 controls failed to meet the validity criteria, corresponding to 39%. Thereby, all invalid controls were derived by L05 (4 out of 10 controls were invalid, i.e. 40%) and L06 (10 out of 14 controls were invalid, i.e. 71%).

For the Bacteria Luminescence test, 9 out of 62 controls were found to be invalid, corresponding to 15%. L06 accounted for most of the invalid controls (7 out of 20, i.e. 35%).

For L03, there was evidence that the LBT tends to detect increased values for toxicity measures compared to other laboratories (mean pseudo z-score LID = 0.972, mean pseudo-z-score %inhibition in D1 = 1.371). The same applies to L05 (LBT, mean pseudo z-score LID 0.872[13]), whereas for L04, the obtained scores point to reduced values for toxicity in the LBT (mean pseudo z-score LID = - 0.691; mean pseudo z-score %inhibition in D1 = -1.441) and increased values of toxicity in the AGI (mean pseudo-z-score LID = 0.606).

[13] The elevated pseudo-z-score for L05 in the LBT may be at least partly due to the fact, that the lab didn´t test D1 and thus LIDs could be only given as $\leq$ 2. These were set to 2 for statistical evaluation, thus, “2” was the lowest LID considered for L05 in the LBT. The resulting pseudo-z-scores thus might be an artefact due to data modification.

## 5. References

[1] Flöter, Carolin (2009): Basisdokument zum Grünalgentest mit *Desmodesmus subspicatus* basierend auf DIN 38 412-L33 (1991) und DIN EN ISO 8692 (2005) für die Bewertung von Sedimenten/ Baggergut im Auftrag der Hamburg Port Authority, Stand 2009_03, Modifiziert im August 2018 durch HPA

[1a] Flöter, Carolin (2009): Basisdokument zur Probenvorbereitung von limnischen Sedimenten für die Durchführung von limnischen Biotestverfahren im Auftrag der Hamburg Port Authority, Stand 2009_03_17.

[2] BfG-Merkblatt "Ökotoxikologische Baggergutuntersuchung", Ökotoxikologische Untersuchung von Sedimenten, Eluaten und Porenwässern, Bundesanstalt für Gewässerkunde, Koblenz: Stand: September 2011

[3a] ISO 5725-1 (1997-11): Accuracy (trueness and precision) of measurement methods and results - Part 1 : General principles and definitions

[3b] ISO 5725-2 (1994): Accuracy (trueness and precision) of measurement methods and results - Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method (lSO 5725-2:1994 including Technical Corrigendum 1:2002)

[4] DIN 38402-45:2003-04: Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung - Allgemeine Angaben (Gruppe A) - Teil 45: Ringversuche zur Eignungsprüfung von Laboratorien (A 45)

[5] Guidance Document on Statistical Methods for Environmental Toxicity Tests. Report EPS 1/RM/46, Environment Canada, 2005

[6] Teuchies, Johannes, and WP3 partners from Sullied sediments project (2017): Sullied sediments – WP3 Sediment assessment Sampling SOP.

[7] European Food Safety Authority, Outcome of the pesticides peer review meeting on general recurring issues in ecotoxicology, TECHNICAL REPORT APPROVED: 18 December 2015 PUBLISHED: 22 December 2015, www.efsa.europa.eu/publications EFSA Supporting publication 2015:EN-924

## Annex

**Annex Table 1a:**
Single laboratory results for the Algae Growth Inhibition test: LID
Red: test invalid; green: blank value slightly elevated, test nevertheless regarded as valid; blue: inhibition of reference slightly elevated, test nevertheless regarded as valid;
Sediment 1,2 and 3: two different samples of the same sediment tested.

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| Porewater | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 2 | 1 | 1 | 8 | 1 | 1 | 2 |
| | 2 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 4 | 1 | 1 | 1 |
| | ≥32 | 1 | 1 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 16 | 16 | 16 | 2 | 1 | 4 | 1 |
| | 16 | 16 | 16 | | | | |
| L5 | 2 | 1 | 4 | 4 | 1 | 2 | 1 |
| | 2 | 4 | 2 | | | | |
| L6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | | | | |

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eluate | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 4 | 2 | 1 | 4 | 1 | 1 | 2 |
| | 2 | 1 | 2 | | | | |
| L2 | 2 | 4 | 1 | 32 | 4 | 8 | 1 |
| | 4 | ≥32 | 1 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | 1 | 16 | ≥32 | 32 | 4 | 1 | 4 |
| | 1 | 16 | 8 | | | | |
| L5 | 2 | 8 | 1 | 4 | 1 | 2 | 1 |
| | 4 | 8 | 1 | | | | |
| L6 | 1 | 8 | 1 | 8 | 1 | 1 | 1 |
| | 1 | 8 | 1 | | | | |

**Annex Table 1b :**
Single laboratory results for the Algae Growth Inhibition test: %inhibition in D1;
Red: test invalid; green: blank value slightly elevated, test nevertheless regarded as valid; blue: inhibition of reference slightly elevated, test nevertheless regarded as valid;
Sediment 1,2 and 3: two different samples of the same sediment tested.

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| Porewater | %inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 42,3 | -32,1 | -85,8 | 85,5 | -204,8 | -2,7 | 57,7 |
| | 29,9 | -87,4 | -92,9 | | | | |
| L2 | 27,0 | -75,0 | -24,0 | 91,0 | -126,0 | -58,0 | 0,0 |
| | 47,0 | -89,0 | -58,0 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | -7,0 | -39,0 | -54,0 | 70,0 | -67,0 | -8,0 | -22,0 |
| | -4,0 | -32,0 | -73,0 | | | | |
| L5 | 20,0 | -1,5 | 32,3 | 64,2 | -31,4 | 52,2 | -6,3 |
| | 41,3 | 29,5 | 27,1 | | | | |
| L6 | -44,8 | -5,0 | -40,7 | -23,0 | -417,7 | -118,2 | -228,4 |
| | -46,7 | 16,8 | -41,8 | | | | |

| Algae Growth Inhibition Test | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eluate | %inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 62,0 | 28,1 | 12,1 | 93,1 | -130,9 | -5,6 | 49,9 |
| | 44,4 | 11,6 | 23,7 | | | | |
| L2 | 80,0 | 89,0 | -26,0 | 66,0 | -142,0 | 4,0 | 10,0 |
| | 74,0 | 81,0 | 6,0 | | | | |
| L3 | - | - | - | - | - | - | - |
| | - | - | - | | | | |
| L4 | -27,0 | 64,0 | 93,0 | 41,0 | 49,0 | -25,0 | 47,0 |
| | -42,0 | 50,0 | 90,0 | | | | |
| L5 | 49,1 | -6,0 | 11,9 | 79,1 | -59,2 | 22,2 | -4,7 |
| | 48,8 | 25,7 | -11,1 | | | | |
| L6 | 30,5 | -67,4 | -108,3 | 25,5 | -135,4 | -35,8 | -44,2 |
| | -42,5 | -63,5 | -63,8 | | | | |

**Annex Table 2a:**
Single laboratory results for the Luminescence Bacteria test: LID;
Red: test invalid; green: sample stored for more than 30 days, test nevertheless regarded as valid;
Sediment 1,2 and 3: two different samples of the same sediment tested.

| Bacteria Luminescence Test DIN EN ISO 11348-2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Porewater | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 1 | 1 | 1 | 4 | 4 | 1 | 1 |
| | 1 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 2 | 2 | 1 | 2 |
| | 4 | 1 | 1 | | | | |
| L3 | 8 | 4 | 2 | ≥16 | 4 | 2 | ≥8 |
| | 8 | 2 | 2 | | | | |
| L4 | 1 | 1 | 1 | 1 | 1 | 2 | 1 |
| | 1 | 1 | 1 | | | | |
| L5 | ≤2 | ≤2 | ≤2 | 4 | 4 | ≤2 | ≤2 |
| | ≤2 | ≤2 | ≤2 | | | | |
| L6 | 2 | 2 | 2 | 1 | 1 | 1 | 1 |
| | 2 | 1 | 2 | | | | |

| Bacteria Luminescence Test DIN EN ISO 11348-2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eluate | LID | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 4 | 1 | 1 | 4 | 1 | 1 | 1 |
| | 4 | 1 | 1 | | | | |
| L2 | 4 | 1 | 1 | 2 | 1 | 1 | 1 |
| | 4 | 1 | 1 | | | | |
| L3 | ≥16 | 1 | 1 | ≥8 | 4 | 2 | 2 |
| | ≥16 | 1 | 1 | | | | |
| L4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 1 | 1 | 1 | | | | |
| L5 | 16 | ≤2 | ≤2 | ≤2 | ≤2 | ≤2 | ≤2 |
| | 8 | ≤2 | ≤2 | | | | |
| L6 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |
| | 2 | 1 | 1 | | | | |

**Annex Table 2b:**
Single laboratory results for the Luminescence Bacteria test: %Inhibition in D1;
Red: test invalid; green: sample stored for more than 30 days, test nevertheless regarded as valid;
Sediment 1,2 and 3: two different samples of the same sediment tested.

| Bacteria Luminescence Test DIN EN ISO 11348-2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Porewater | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 16,2 | 3,5 | 4,6 | 27,0 | 32,9 | -3,9 | 11,4 |
| | 15,2 | 1,8 | 8,3 | | | | |
| L2 | 42,3 | 17,6 | 13,9 | 32,0 | 27,7 | 12,1 | 28,4 |
| | 47,6 | 17,7 | 15,2 | | | | |
| L3 | 57,1 | 31,8 | 31,2 | 87,4 | 72,2 | 41,8 | 62,9 |
| | 58,7 | 33,1 | 28,3 | | | | |
| L4 | -78,9 | -32,8 | -34,4 | -72,9 | -24,0 | 57,7 | 10,9 |
| | -65,1 | -70,0 | -72,9 | | | | |
| L5 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| | n.d. | n.d. | n.d. | | | | |
| L6 | 31,0 | 27,5 | 23,4 | 4,6 | 1,3 | -25,7 | 18,8 |
| | 29,9 | 18,0 | 10,3 | | | | |

| Bacteria Luminescence Test DIN EN ISO 11348-2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Eluate | % Inhibition in D1 | | | | | | |
| | Sed 1 | Sed 2 | Sed 3 | Sed 4 | Sed 5 | Sed 6 | Sed 7 |
| L1 | 36,2 | 1,8 | -0,01 | 33,2 | -5,2 | -6,2 | 1,1 |
| | 48,7 | -2,4 | -2,6 | | | | |
| L2 | 43,6 | 15,1 | 2,9 | 48,5 | 7,6 | 12,9 | 5,4 |
| | 43,4 | 5,5 | 6,8 | | | | |
| L3 | 76,8 | 15,4 | 15,7 | 66,6 | 57,6 | 46,5 | 50,2 |
| | 86,8 | 16,4 | 10,8 | | | | |
| L4 | -14,9 | -55,4 | -64,7 | -13,2 | -36,3 | -49,3 | -49,5 |
| | -38,4 | -62,4 | -69,5 | | | | |
| L5 | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| | n.d. | n.d. | n.d. | | | | |
| L6 | -6,4 | 3,5 | -3,3 | 43,1 | 3,9 | 0,9 | 1,4 |
| | 36,4 | -5,8 | -5,6 | | | | |

**Annex Table 3: Results of statistical evaluation of LIDs obtained in the Algae Growth Inhibition test performed with eluate samples.**
Sediment 1,2 and 3: two different samples of the same sediment tested.
* all calculations with ln-transformed data, geomean and sR by retransformation of the parameters of the log-normal distribution, except for Sr and sL. red: "nearly outlier" according to warningchart concept; i.e. outside the 95% PI
§ LID < x and > x were set to x
# Sediment 5 and 7 (italics): no invalid tests, hence same results as based on all data.

| | **Algae Growth Inhibition Test, LID, Eluate** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 3 | 3 | 2 | 1 | 0 | 1 | 0 |
| % invalid measurements | 30,0 | 30,0 | 20,0 | 20,0 | 0,0 | 20,0 | 0,0 |
| | **Laboratory specific LID, including invalid data §** | | | | | | |
| L01 | 4 - 2 | 2 - 1 | 1 - 2 | 4 | 1 | 1 | 2 |
| L02 | 2 - 4 | 4 - 32 | 1 - 1 | 32 | 4 | 8 | 1 |
| L03 | test not performed | | | | | | |
| L04 | 1 - 1 | 16 - 16 | 32 - 8 | 32 | 4 | 1 | 4 |
| L05 | 2 - 4 | 8 - 8 | 1 - 1 | 4 | 1 | 2 | 1 |
| L06 | 1 - 1 | 8 - 8 | 1 - 1 | 8 | 1 | 1 | 1 |
| min / max | 1 / 4 | 1 / 32 | 1 / 32 | 4 / 32 | 1 / 4 | 1 / 8 | 1 / 4 |
| | **LID results including invalid data *** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **geometric mean** | **1,87** | **6,97** | **1,87** | **10,60** | **1,74** | **1,74** | **1,52** |
| 95% CI | 0,6-4,1 | 1,9-25,9 | 0,4-8,9 | 2,9-38,9 | 0,7-4,5 | 0,6-5,3 | 0,7-3,3 |
| 95% PI | 0,5-6,4 | 0,9-55,3 | 0,2-22,0 | 1,3-82,9 | 0,4-7,7 | 0,3-10,2 | 0,5-5,1 |
| NW (95%CI / mean) | 1.9 | 3.4 | 4.5 | 3.4 | 2.2 | 2.7 | 1.7 |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 1,6 | 17,5 | 8,1 | 26,1 | 2,1 | 2,9 | 1,3 |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **69,7** | **143,3** | **196,9** | **142,1** | **88,2** | **112,3** | **68,4** |
| | **Laboratory specific LID, only valid data §#** | | | | | | |
| L01 | 4 - 2 | 2 - 1 | 1 - 2 | 4 | *1* | 1 | *2* |
| L02 | 2 - 4 | 4 - 32 | 1 - 1 | 32 | *4* | 8 | *1* |
| L03 | test not performed | | | | | | |
| L04 | 1 - 1 | 16 - 16 | 32 - 8 | 32 | *4* | 1 | *4* |
| L05 | - 4 | - 8 | 1 - 1 | 4 | *1* | 2 | *1* |
| L06 | - - | - - | - - | - | *1* | - | *1* |
| min / max | 1 / 4 | 1 / 32 | 1 / 32 | 4 / 32 | *1 / 4* | 1 / 8 | *1 / 4* |
| | **LID results only valid data * #** | | | | | | |
| number of laboratories for statistics | 4 | 4 | 4 | 4 | *5* | 4 | *5* |
| number of measurements for statistics | 7 | 7 | 8 | 4 | *5* | 4 | *5* |
| **geometric mean** | **2,21** | **6,56** | **2,2** | **11,31** | ***1,74*** | **2,00** | ***1,52*** |
| 95% CI | 0,8-6,2 | 0,8-51,7 | 0,2-20,0 | 1,7-76,4 | *0,7-4,5* | 0,4 - 9,5 | *0,7-3,3* |
| 95% PI | 0,6-7,9 | 0,5-83,3 | 0,1-33,5 | 1,1-119,0 | *0,4-7,7* | 0,3-13,7 | *0,5-5,1* |
| NW (95%CI / mean) | 2.9 | 7.3 | 10.6 | 7.0 | *2.2* | 5.2 | *1.7* |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 2,0 | 31,8 | 14,1 | 41,8 | *2,1* | 4,1 | *1,3* |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **72,9** | **209,1** | **244,2** | **179,6** | ***88,2*** | **127,0** | ***68,4*** |

**Annex Table 4: Results of statistical evaluation of LIDs obtained in the Algae Growth Inhibition test performed with pore water samples.**
Sediment 1,2 and 3: two different samples of the same sediment tested.
* all calculations with ln-transformed data, geomean and sR by retransformation of the parameters of the log-normal distribution, except for Sr and sL.
§ LID < x and > x were set to x

| | **Algae Growth Inhibition Test, LID, Pore Water** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 3 | 2 | 3 | 1 | 1 | 2 | 2 |
| % invalid measurements | 30,0 | 20,0 | 30,0 | 20,0 | 20,0 | 40,0 | 40,0 |
| | **Laboratory specific LID, including invalid data §** | | | | | | |
| L01 | 2 - 2 | 1 - 1 | 1 - 1 | 8 | 1 | 1 | 2 |
| L02 | 4 - 32 | 1 - 1 | 1 - 1 | 4 | 1 | 1 | 1 |
| L03 | test not performed | | | | | | |
| L04 | 16 - 16 | 16 - 16 | 16 - 16 | 2 | 1 | 4 | 1 |
| L05 | 2 - 2 | 1 - 4 | 4 - 2 | 4 | 1 | 2 | 1 |
| L06 | 1 - 1 | 1 - 1 | 1 - 1 | 1 | 1 | 1 | 1 |
| min / max | 1 / 32 | 1 / 16 | 1 / 16 | 1 / 8 | 1 / 1 | 1 / 4 | 1 / 2 |
| | **LID results including invalid data *** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **geometric mean** | **3,73** | **2** | **2,14** | **3,03** | **1,00** | **1,52** | **1,21** |
| 95% CI | 0,7-18,7 | 0,4-9,3 | 0,5-9,7 | 1,1-8,1 | n.d. | 0,7-3,3 | 0,5-1,7 |
| 95% PI | 0,3-47,4 | 0,2-22,7 | 0,2-23,4 | 0,6-14,3 | n.d. | 0,5-5,1 | 0,6-2,1 |
| NW (95%CI / mean) | 4.8 | 4.5 | 4.3 | 2.3 | - | 1.7 | 1.0 |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 18,1 | 8,2 | 8,3 | 3,9 | 0,0 | 1,3 | 0,4 |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **209,1** | **191,0** | **185,20** | **93,10** | **0,00** | **68,40** | **31,70** |
| | **Laboratory specific LID, only valid data §** | | | | | | |
| L01 | 2 - 2 | 1 - 1 | 1 - 1 | 8 | 1 | 1 | 2 |
| L02 | 4 - 32 | 1 - 1 | 1 - 1 | 4 | 1 | 1 | 1 |
| L03 | test not performed | | | | | | |
| L04 | 16 - 16 | 16 - 16 | 16 - 16 | 2 | 1 | 4 | 1 |
| L05 | - 2 | - - | 4 - | 4 | 1 | - | - |
| L06 | - - | 1 - 4 | - - | - | - | - | - |
| min / max | 2 / 32 | 1 / 16 | 1 / 16 | 1 / 8 | 1 / 1 | 1 / 4 | 1 / 2 |
| | **LID results only valid data *** | | | | | | |
| number of laboratories for statistics | 4 | 4 | 4 | 4 | 4 | 3 | 3 |
| number of measurements for statistics | 7 | 8 | 7 | 4 | 4 | 3 | 3 |
| **geometric mean** | **5,94** | **2** | **2,7** | **4,00** | **1,00** | **1,59** | **1,26** |
| 95% CI | 0,8-42,7 | 0,2-18,2 | 0,3-26,0 | 1,6-9,8 | n.d. | 0,2-11,6 | 0,5-3,4 |
| 95% PI | 0,5-67,4 | 0,1-30,3 | 0,2-43,9 | 1,3-12,1 | n.d. | 0,3-7,6 | 0,6-2,8 |
| NW (95%CI / mean) | 7.1 | 9.0 | 9.6 | 2.1 | - | 7.2 | 2.3 |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 24,4 | 12,6 | 19,1 | 2,9 | 0,0 | 2,1 | 0,6 |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **190,8** | **241,3** | **256,7** | **61,4** | **0,0** | **94,7** | **41,6** |

**Annex Table 5: Results of statistical evaluation of %Inhibition in D1 obtained in the Algae Growth Inhibition test performed with eluate samples.** Sediment 1,2 and 3: two different samples of the same sediment tested.

* sr can only be calculated if repeated measurements were performed, i.d. for Sed 1, 2 and 3

§ inhibitions < 0 (i.e. stimulations) were set to 0% inhibition

# Sediment 5 and 7 (in italics): no invalid tests, hence same results as based on all data.

| | **Algae Growth Inhibition test, % Inhibition in D1, Eluate** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 3 | 3 | 2 | 1 | 0 | 1 | 0 |
| % invalid measurements | 30,0 | 30,0 | 20,0 | 20,0 | 0,0 | 20,0 | 0,0 |
| | **Laboratory specific % Inhibition in D1, including invalid data §** | | | | | | |
| L01 | 62,0 / 44,4 | 28,1 / 11,6 | 12,1 / 23,7 | 93,1 | 0 | 0 | 49,9 |
| L02 | 80,0 / 74,0 | 89,0 / 81,0 | 0 / 6 | 66,0 | 0 | 4,0 | 10,0 |
| L03 | test not performed | | | | | | |
| L04 | 0 / 0 | 64,0 / 50,0 | 93,0 / 90,0 | 41,0 | 49,0 | 0 | 47,0 |
| L05 | 49,1 / 48,8 | 0 / 25,7 | 11,9 / 0 | 79,5 | 0 | 22,2 | 0 |
| L06 | 30,5 / 0 | 0 / 0 | 0 / 0 | 25,5 | 0 | 0 | 0 |
| min / max | 0 / 80,0 | 0 / 89,0 | 0 / 93,0 | 25,5 / 93,1 | 0 / 49,0 | 0 / 22,2 | 0 / 49,9 |
| | **% Inhibition in D1, results including invalid data** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **mean** | **38,9** | **34,9** | **23,7** | **61,0** | **9,8** | **5,2** | **21,4** |
| 95% CI | 0 -78,6 | 0 -79,6 | 0 -71,8 | 26,7-95,4 | 0 -37,0 | 0 - 17,2 | 0 - 52,5 |
| 95% PI | 0 -101,5 | 0 -105,4 | 0 -99,6 | 6,8-115,2 | 0 - 52,8 | 0 - 24,1 | 0 - 70,5 |
| NW (95%CI / mean) | 2-0 | 2.3 | 3.0 | 1.1 | 3.8 | 4.4 | 2.4 |
| sr (repeatability)* | 11,3 | 10,9 | 5,7 | n.d. | n.d. | n.d. | n.d. |
| sL (inter-laboratory variability) | 29,9 | 34,2 | 38,3 | 27,6 | 21,9 | 9,6 | 25,1 |
| sR (reproducibility) | 32,0 | 35,9 | 38,7 | 27,6 | 21,9 | 9,6 | 25,1 |
| sr % | 29,1 | 31,3 | 23,9 | n.d. | n.d. | n.d. | n.d. |
| sL % | 76,9 | 98,0 | 161,9 | n.d. | n.d. | n.d. | n.d. |
| **sR %** | **82,2** | **102,9** | **163,6** | **45,3** | **223,5** | **183,9** | **117,2** |
| | **Laboratory specific % Inhibition in D1, only valid data § #** | | | | | | |
| L01 | 62,0 / 44,4 | 28,1 / 11,6 | 12,1 / 23,7 | 93,1 | *0* | 0 | *49,9* |
| L02 | 80,0 / 74,0 | 89,0 / 81,0 | 0 / 6 | 66,0 | *0* | 4,0 | *10,0* |
| L03 | test not performed | | | | | | |
| L04 | 0 / 0 | 64,0 / 50,0 | 93,0 / 90,0 | 41,0 | *49,0* | 0 | *47,0* |
| L05 | - / 48,8 | - / 25,7 | 11,9 / 0 | 79,5 | *0* | 22,2 | *0* |
| L06 | - / - | - / - | - / - | - | *0* | 0 | *0* |
| min / max | 0 / 80,0 | 11,6 / 89,0 | 0 / 93,0 | 41,0 / 93,1 | *0 / 49,0* | 0 / 22,2 | *0 / 49,9* |
| | **% Inhibition in D1 - results only valid data #** | | | | | | |
| number of laboratories for statistics | 4 | 4 | 4 | 4 | *5* | 4 | *5* |
| number of measurements for statistics | 7 | 7 | 8 | 4 | *5* | 4 | *5* |
| **mean** | **44,20** | **49,9** | **29,6** | **69,90** | ***9,8*** | **6,6** | ***21,4*** |
| 95% CI | 0 - 100,1 | 0 - 100,3 | 0 - 96,4 | 34,5 - 105,3 | *0 -37,0* | 0 - 23,4 | *0 - 52,5* |
| 95% PI | 0 - 113,1 | 0 - 111,9 | 0 - 111,9 | 26,3 - 113,4 | *0 - 52,8* | 0 - 27,3 | *0 - 70,5* |
| NW (95%CI / mean) | 2.3 | 2.0 | 3.3 | 1.0 | *3.8* | 3.5 | *2.4* |
| sr (repeatability)* | 7,8 | 9,4 | 6,3 | n.d. | *n.d.* | n.d. | *n.d.* |
| sL (inter-laboratory variability) | 34,4 | 30,2 | 41,5 | 22,2 | *21,9* | 10,6 | *25,1* |
| sR (reproducibility) | 35,2 | 31,6 | 42,0 | 22,2 | *21,9* | 10,6 | *25,1* |
| sr % | 17,2 | 18,9 | 21,4 | n.d. | *n.d.* | n.d. | *n.d.* |
| sL % | 77,8 | 60,5 | 140,4 | n.d. | *n.d.* | n.d. | *n.d.* |
| **sR %** | **79,6** | **63,4** | **142,0** | **31,8** | ***223,5*** | **161,9** | ***117,2*** |

**Annex Table 6: Results of statistical evaluation of %Inhibition in D1 obtained in the Algae Growth Inhibition test performed with pore water samples.** Sediment 1,2 and 3: two different samples of the same sediment tested.
red: "nearly outliers" according to warningchart concept; i.e. outside the 95% PI
* sr can only be calculated if repeated measurements were performed, i.d. for Sed 1, 2 and 3
§ inhibitions < 0 (i.e. stimulations) were set to 0% inhibition
a artefacts due to data situation: sr > sR; sL <0

| | **Algae Growth Inhibition test, % Inhibition in D1 - Pore Water** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 3 | 2 | 3 | 1 | 1 | 2 | 2 |
| % invalid measurements | 30,0 | 20,0 | 30,0 | 20,0 | 20,0 | 40,0 | 40,0 |
| | **Laboratory specific % Inhibition in D1, including invalid data §** | | | | | | |
| L01 | 42,3 / 29,9 | 0 / 0 | 0 / 0 | 85,5 | 0 | 0 | 57,7 |
| L02 | 27,0 / 47,0 | 0 / 0 | 0 / 0 | 91,0 | 0 | 0 | 0 |
| L03 | test not performed | | | | | | |
| L04 | 0 / 0 | 0 / 0 | 0 / 0 | 70,0 | 0 | 0 | 0 |
| L05 | 20,0 / 41,3 | 0 / 29,5 | 32,3 / 27,1 | 64,2 | 0 | 52,2 | 0 |
| L06 | 0 / 0 | 0 / 16,8 | 0 / 0 | 0 | 0 | 0 | 0 |
| min / max | 0 / 47,0 | 0 / 29,5 | 0 / 27,1 | 0 / 91,0 | 0 / 0 | 0 / 52,2 | 0 / 57,7 |
| | **% Inhibition in D1, results including invalid data** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **mean** | **20,8** | **4,6** | **5,9** | **62,1** | **0,0** | **10,4** | **11,5** |
| 95% CI | 0 - 46,0 | 0 - 17,2 | 0 - 22,5 | 16,9 - 107,4 | n.d. | 0 - 39,4 | 0 - 43,6 |
| 95% PI | 0 - 60,7 | 0 - 24,5 | 0 - 32,0 | 0 - 133,5 | n.d. | 0 - 56,2 | 0 - 62,1 |
| NW (95%CI / mean) | 2.2 | 3.7 | 3.8 | 1.5 | - | 3.8 | 3.8 |
| sr (repeatability)* | 10,0 | 10,7 a | 1,6 | n.d. | n.d. | n.d. | n.d. |
| sL (inter-laboratory variability) | 17,7 | n.d. a | 13,2 | 36,4 | 0,0 | 23,3 | 25,8 |
| sR (reproducibility) | 20,4 | 10,1 a | 13,3 | 36,4 | 0,0 | 23,3 | 25,8 |
| sr % | 48,4 | 231,9 a | 27,7 | n.d. | n.d. | n.d. | n.d. |
| sL % | 85,4 | n.d. a | 222,7 | n.d. | 0,0 | n.d. | n.d. |
| **sR %** | **98,2** | **219,0** | **224,5** | **58,6** | **0,0** | **223,6** | **223,6** |
| | **Laboratory specific % Inhibition in D1, only valid data §** | | | | | | |
| L01 | 42,3 / 29,9 | 0 / 0 | 0 / 0 | 85,5 | 0 | 0 | 57,7 |
| L02 | 27,0 / 47,0 | 0 / 0 | 0 / 0 | 91,0 | 0 | 0 | 0 |
| L03 | test not performed | | | | | | |
| L04 | 0 / 0 | 0 / 0 | 0 / 0 | 70,0 | 0 | 0 | 0 |
| L05 | - / 41,3 | - / - | 32,3 / - | 64,2 | 0 | - | - |
| L06 | - / - | 0 / 16,8 | - / - | - | 0 | - | - |
| min / max | 0 / 47,0 | 0 / 16,8 | 0 / 32,3 | 64,2 / 91,0 | 0 / 0 | 0 / 0 | 0 / 57,7 |
| | **% Inhibition in D1 - results only valid data** | | | | | | |
| number of laboratories for statistics | 4 | 4 | 4 | 4 | 5 | 3 | 3 |
| number of measurements for statistics | 7 | 8 | 7 | 4 | 5 | 3 | 3 |
| **mean** | **26,8** | **2,1** | **4,6** | **77,7** | **0,0** | **0,0** | **19,2** |
| 95% CI | 0 - 59,9 | 0 - 11,6 | 0 - 25,6 | 57,6 - 97,8 | n.d. | n.d. | 0 - 102,0 |
| 95% PI | 0 - 67,6 | 0 - 13,7 | 0 - 30,5 | 52,9 - 102,4 | n.d. | n.d. | 0 - 84,0 |
| NW (95%CI / mean) | 2.2 | 5.5 | 5.6 | 0.5 | n.d. | n.d. | 5.3 |
| sr (repeatability)* | 9,6 | 5,9 | 0,0 | n.d. | n.d. | n.d. | n.d. |
| sL (inter-laboratory variability) | 18,5 | 0,0 | 13,2 | 12,6 | 0,0 | 0,0 | 33,3 |
| sR (reproducibility) | 20,8 | 5,9 | 13,2 | 12,6 | 0,0 | 0,0 | 33,3 |
| sr % | 35,9 | 282,8 | 0,0 | n.d. | n.d. | n.d. | n.d. |
| sL % | 68,9 | 0,0 | 285,8 | 16,3 | 0,0 | 0,0 | 173,2 |
| **sR %** | **77,7** | **282,8** | **285,8** | **16,3** | **0,0** | **0,0** | **173,2** |

**Annex Table 7: Results of statistical evaluation of LIDs obtained in the Luminescence Bacteria test performed with eluate samples.**
Sediment 1,2 and 3: two different samples of the same sediment tested.
* all calculations with ln-transformed data, geomean and sR by retransformation of the parameters of the log-normal distribution, except for Sr and sL.
§ LID < x and > x were set to x
# Sediment 2, 4, 6 and 7 (italics): no invalid tests, hence same results as based on all data.

| | **Bacteria Luminescence Test, LID, Eluate** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| number of measurements | 12 | 12 | 12 | 6 | 6 | 6 | 6 |
| invalid measurements | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| % invalid measurements | 8,3 | 0,0 | 8,3 | 0,0 | 16,6 | 0,0 | 0,0 |
| | **Laboratory specific LID, including invalid data §** | | | | | | |
| L01 | 4 - 4 | 1 - 1 | 1 - 1 | 4 | 1 | 1 | 1 |
| L02 | 4 - 4 | 1 - 1 | 1 - 1 | 2 | 1 | 1 | 1 |
| L03 | 16 - 16 | 1 - 1 | 1 - 1 | 8 | 4 | 2 | 2 |
| L04 | 1 - 1 | 1 - 1 | 1 - 1 | 1 | 1 | 1 | 1 |
| L05 | 16 - 8 | 2 - 2 | 2 - 2 | 2 | 2 | 2 | 2 |
| L06 | 1 - 2 | 1 - 1 | 1 - 1 | 2 | 1 | 1 | 1 |
| min / max | 1 - 16 | 1 - 2 | 1 - 2 | 1 - 8 | 1 - 4 | 1 - 2 | 1 - 2 |
| | **LID results including invalid data *** | | | | | | |
| number of laboratories for statistics | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| number of measurements for statistics | 12 | 12 | 12 | 6 | 6 | 6 | 6 |
| **geometric mean** | **4,00** | **1,12** | **1,12** | **2,52** | **1,42** | **1,26** | **1,26** |
| 95% CI | 1,2 - 12,9 | 0,8 - 1,5 | 0,8 - 1,5 | 1,2 - 5,3 | 0,8 - 2,6 | 0,9 - 1,8 | 0,9 - 1,8 |
| 95% PI | 0,5 - 35,5 | 0,6 - 2,0 | 0,6 - 2,0 | 0,6 - 10,2 | 0,5 - 4,4 | 0,6 - 2,5 | 0,6 - 2,5 |
| NW (95%CI / mean) | 2.9 | 0.6 | 0.6 | 1.6 | 1.3 | 0.7 | 0.7 |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 11,7 | 0,3 | 0,3 | 2,7 | 1,1 | 0,5 | 0,5 |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **156,7** | **28,8** | **28,8** | **81,8** | **63,2** | **36,9** | **36,9** |
| | **Laboratory specific LID, only valid data § #** | | | | | | |
| L01 | 4 - 4 | *1 - 1* | 1 - 1 | *4* | 1 | *1* | *1* |
| L02 | 4 - 4 | *1 - 1* | 1 - 1 | *2* | 1 | *1* | *1* |
| L03 | 16 - 16 | *1 - 1* | 1 - 1 | *8* | 4 | *2* | *2* |
| L04 | 1 - | *1 - 1* | 1 - 1 | *1* | 1 | *1* | *1* |
| L05 | 16 - 8 | *2 - 2* | 2 - | *2* | 2 | *2* | *2* |
| L06 | 1 - 2 | *1 - 1* | 1 - | *2* | - | *1* | *1* |
| min / max | 1 - 16 | *1 - 2* | 1 - 2 | *1 - 8* | 1 - 4 | *1 - 2* | *1 - 2* |
| | **LID results only valid data * #** | | | | | | |
| number of laboratories for statistics | 6 | *6* | 6 | *6* | 5 | *6* | *6* |
| number of measurements for statistics | 11 | *12* | 11 | *6* | 5 | *6* | *6* |
| **geometric mean** | **4,53** | ***1,12*** | **1,1** | ***2,52*** | **1,52** | ***1,26*** | ***1,26*** |
| 95% CI | 1,5 - 13,9 | *0,8 - 1,5* | 0,8 - 1,5 | *1,2 - 5,3* | 0,7 - 3,3 | *0,9 - 1,8* | *0,9 - 1,8* |
| 95% PI | 0,6 - 36,6 | *0,6 - 2,0* | 0,6 - 2,0 | *0,6 - 10,2* | 0,5 - 5,1 | *0,6 - 2,5* | *0,6 - 2,5* |
| NW (95%CI / mean) | 2.7 | *0.6* | 0.6 | *1.6* | 1.2 | *0.7* | *0.7* |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 11,6 | *0,3* | 0,4 | *2,7* | 1,3 | *0,5* | *0,5* |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **145,1** | ***28,8*** | **30,0** | ***81,8*** | **68,4** | ***36,9*** | ***36,9*** |

**Annex Table 8: Results of statistical evaluation of LIDs obtained in the Bacteria Luminescence test performed with pore water samples.**
Sediment 1,2 and 3: two different samples of the same sediment tested.
Red: "nearly outlier" according to warningchart concept; i.e. outside the 95% PI
* all calculations with ln-transformed data, geomean and sR by retransformation of the parameters of the log-normal distribution, except for Sr and sL.
§ LID < x and > x were set to x
# Sediment 1, 5 and 7 (italics): no invalid tests, hence same results as based on all data;

| | **Bacteria Luminescence Test, LID, Pore Water** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| number of measurements | 12 | 12 | 12 | 6 | 6 | 6 | 6 |
| invalid measurements | 0 | 2 | 1 | 1 | 0 | 2 | 1 |
| % invalid measurements | 0,0 | 16,6 | 8,3 | 16,6 | 0,0 | 33,3 | 0,0 |
| | **Laboratory specific LID, including invalid data §** | | | | | | |
| L01 | 1 - 1 | 1 - 1 | 1 - 1 | 4 | 4 | 1 | 1 |
| L02 | 4 - 4 | 1 - 1 | 1 - 1 | 2 | 2 | 1 | 2 |
| L03 | 8 - 8 | 4 - 2 | 2 - 2 | 16 | 4 | 2 | 8 |
| L04 | 1 - 1 | 1 - 1 | 1 - 1 | 1 | 1 | 2 | 1 |
| L05 | 2 - 2 | 2 - 2 | 2 - 2 | 4 | 4 | 2 | 2 |
| L06 | 2 - 2 | 2 - 1 | 2 - 2 | 1 | 1 | 1 | 1 |
| min / max | 1 - 8 | 1 - 4 | 1 - 2 | 1 - 16 | 1 - 4 | 1 - 2 | 1 - 8 |
| | **LID results including invalid data *** | | | | | | |
| number of laboratories for statistics | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| number of measurements for statistics | 12 | 12 | 12 | 6 | 6 | 6 | 6 |
| **geometric mean** | **2,25** | **1,42** | **1,42** | **2,83** | **2,45** | **1,41** | **1,78** |
| **95% CI** | 1,0 - 5,3 | 0,9 - 2,3 | 1,0 - 2,1 | 0,9 - 8,5 | 1,1 - 4,6 | 1,0 - 2,1 | 0,7 - 4,2 |
| 95% PI | 0,5 - 11,0 | 0,6 - 3,6 | 0,7 - 3,0 | 0,4 - 22,2 | 0,6 - 8,5 | 0,7 - 3,0 | 0,4 - 8,7 |
| NW (95%CI / mean) | 1.9 | 1.0 | 0.8 | 2.7 | 1.4 | 0.8 | 2.0 |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | 3,0 | 0,8 | 0,6 | 7,0 | 2,2 | 0,6 | 2,4 |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | **96,3** | **51,1** | **39,3** | **142,0** | **76,8** | **39,3** | **96,3** |
| | **Laboratory specific LID, only valid data § #** | | | | | | |
| L01 | *1 - 1* | 1 - 1 | 1 - 1 | 4 | *4* | 1 | *1* |
| L02 | *4 - 4* | 1 - 1 | 1 - 1 | 2 | *2* | 1 | *2* |
| L03 | *8 - 8* | 4 - 2 | 2 - 2 | 16 | *4* | 2 | *8* |
| L04 | *1 - 1* | 1 - 1 | 1 - 1 | 1 | *1* | 2 | *1* |
| L05 | *2 - 2* | 2 - 2 | 2 - 2 | 4 | *4* | - | *2* |
| L06 | *2 - 2* | - - | - 2 | - | *1* | - | *1* |
| min / max | *1 - 8* | 1 - 4 | 1 - 2 | 1 - 16 | *1 - 4* | 1 - 2 | *1 - 8* |
| | **LID results only valid data * #** | | | | | | |
| number of laboratories for statistics | *6* | 5 | 6 | 5 | *6* | 4 | *6* |
| number of measurements for statistics | *12* | 10 | 11 | 5 | *6* | 4 | *6* |
| **geometric mean** | ***2,25*** | **1,42** | **1,37** | **3,48** | ***2,45*** | **1,41** | ***1,78*** |
| 95% CI | *1,0 - 5,3* | 0,7 - 2,7 | 0,9 - 2,0 | 1,0 - 12,5 | *1,1 - 4,6* | 0,7 - 2,7 | *0,7 - 4,2* |
| 95% PI | *0,5 - 11,0* | 0,5 - 3,9 | 0,7 - 2,9 | 0,5 - 26,1 | *0,6 - 8,5* | 0,6 - 3,1 | *0,4 - 8,7* |
| NW (95%CI / mean) | *4.3* | 2.0 | 1.0 | 11.5 | *3.5* | 1.1 | *3.5* |
| sr (repeatability)* | - | - | - | - | - | - | - |
| sL (inter-laboratory variability)* | - | - | - | - | - | - | - |
| sR (reproducibility) | *3,0* | 0,9 | 0,6 | 8,1 | *2,2* | 0,6 | *2,4* |
| sr % | - | - | - | - | - | - | - |
| sL % | - | - | - | - | - | - | - |
| **sR %** | ***96,3*** | **54,9** | **39,3** | **137,0** | ***76,8*** | **41,6** | ***96,3*** |

**Annex Table 9: Results of statistical evaluation of %Inhibition in D1 obtained in the Bacteria Luminescence test performed with eluate samples.** Sediment 1,2 and 3: two different samples of the same sediment tested.
Red: "nearly ouitlier" according to warningchart concept; i.e. outside the 95% PI
* sr can only be calculated if repeated measurements were performed, i.e. for Sed 1, 2 and 3
§ inhibitions < 0 (i.e. stimulations) were set to 0% inhibition
a L5 didn´t test D1, therefore no % inhibitions for D1 available
# Sediment 2,4,6 and 7 (in italics): no invalid tests, hence same results as based on all data

| | **Bacteria Luminescence Testm, % Inhibition in D1, Eluate** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| Eluate (EL) or Pore Water (PW) | **EL** | **EL** | **EL** | **EL** | **EL** | **EL** | **EL** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| % invalid measurements | 10,0 | 0,0 | 10,0 | 0,0 | 20,0 | 0,0 | 0,0 |
| | **Laboratory specific % inhibition in D1, including invalid data §** | | | | | | |
| L01 | 36,2 - 48,7 | 1,8 - 0 | 0 - 0 | 33,2 | 0 | 0 | 1,1 |
| L02 | 43,6 - 43,4 | 15,1 - 5,5 | 2,9 - 6,8 | 48,5 | 7,6 | 12,9 | 5,4 |
| L03 | 76,8 - 86,8 | 15,4 - 16,4 | 15,7 - 10,8 | 66,6 | 57,6 | 46,5 | 50,2 |
| L04 | 0 - 0 | 0 - 0 | 0 - 0 | 0 | 0 | 0 | 0 |
| L05 a | - | - | - | - | - | - | - |
| L06 | 0 - 36,4 | 3,5 - 0 | 0 - 0 | 43,1 | 3,9 | 0,9 | 1,4 |
| min / max | 0 - 86,8 | 0 - 16,4 | 0 - 15,7 | 0 - 66,6 | 0 - 57,6 | 0 - 46,5 | 0 - 50,2 |
| | **% inhibition in D1, results, including invalid data** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **mean** | **37,2** | **5,8** | **3,6** | **38,3** | **13,8** | **12,1** | **11,6** |
| 95% CI | 0 - 77,0 | 0 - 14,9 | 0 - 11,0 | 7,9 - 68,8 | 0 - 44,5 | 0 - 36,9 | 0 - 38,5 |
| 95% PI | 0 - 100,1 | 0 - 20,2 | 0 - 15,3 | 0 - 86,5 | 0 - 62,2 | 0 - 51,3 | 0 - 54,1 |
| NW (95%CI / mean) | 2.1 | 2.6 | 3.1 | 1.6 | 3.2 | 3.0 | 3.3 |
| sr (repeatability)* | 12,6 | 3,3 | 2,0 | n.d. | n.d. | n.d. | n.d. |
| sL (inter-laboratory variability) | 29,5 | 6,6 | 5,6 | 24,6 | 24,7 | 20,0 | 21,7 |
| sR (reproducibility) | 32,1 | 7,4 | 6,0 | 24,6 | 24,7 | 20,0 | 21,7 |
| sr % | 33,8 | 57,1 | 54,7 | n.d. | n.d. | n.d. | n.d. |
| sL % | 79,4 | 114,5 | 154,9 | 64,3 | 178,6 | 166,0 | 186,4 |
| **sR %** | **86,3** | **128,0** | **164,2** | **64,3** | **178,6** | **166,0** | **186,4** |
| | **Laboratory specific % inhibition in D1, only valid data § #** | | | | | | |
| L01 | 36,2 - 48,7 | *1,8 - 0* | 0 - 0 | *33,2* | 0 | *0* | *1,1* |
| L02 | 43,6 - 43,4 | *15,1 - 5,5* | 2,9 - 6,8 | *48,5* | 7,6 | *12,9* | *5,4* |
| L03 | 76,8 - 86,8 | *15,4 - 16,4* | 15,7 - 10,8 | *66,6* | 57,6 | *46,5* | *50,2* |
| L04 | 0 - | *0 - 0* | 0 - 0 | *0* | 0 | *0* | *0* |
| L05 a | - | - | - | - | - | - | - |
| L06 | 0 - 36,4 | *3,5 - 0* | 0 - | *43,1* | | *0,9* | *1,4* |
| min / max | | *0 - 16,4* | | *0 - 66,6* | | *0 - 46,5* | *0 - 50,2* |
| | **% inhibition in D1, results, only valid data #** | | | | | | |
| number of laboratories for statistics | 5 | *5* | 5 | *5* | 4 | *5* | *5* |
| number of measurements for statistics | 9 | *10* | 9 | *5* | 4 | *5* | *5* |
| **mean** | **41,3** | ***5,8*** | **4,0** | ***38,3*** | **16,3** | ***12,1*** | ***11,6*** |
| 95% CI | 3,3 - 79,4 | *0 - 14,9* | 0 - 11,6 | *7,9 - 68,8* | 0 - 60,5 | *0 - 36,9* | *0 - 38,5* |
| 95% PI | 0 - 101,4 | *0 - 20,2* | 0 - 16,1 | *0 - 86,5* | 0 - 70,7 | *0 - 51,3* | *0 - 54,1* |
| NW (95%CI / mean) | 1.8 | *2.6* | 2.9 | *1.6* | 3.7 | *3.0* | *3.3* |
| sr (repeatability)* | 14,1 | *3,3* | 2,2 | *n.d.* | n.d. | *n.d.* | *n.d.* |
| sL (inter-laboratory variability) | 27,2 | *6,6* | 5,7 | *24,6* | 27,80 | *20,0* | *21,7* |
| sR (reproducibility) | 30,7 | *7,4* | 6,1 | *24,6* | 27,8 | *20,0* | *21,7* |
| sr % | 34,0 | *57,1* | 55,0 | *n.d.* | n.d. | *n.d.* | *n.d.* |
| sL % | 65,9 | *114,5* | 142,3 | *64,3* | 170,30 | *166,0* | *186,4* |
| **sR %** | **74,2** | ***128,0*** | **152,6** | ***64,3*** | **170,3** | ***166,0*** | ***186,4*** |

Annex Table 10: Results of statistical evaluation of %Inhibition in D1 obtained in the bacteria Luminescence test performed with pore water samples.

Sediment 1,2 and 3: two different samples of the same sediment tested. Red: "nearly ouitlier" according to warningchart concept; i.e. outside the 95% PI

* sr can only be calculated if repeated measurements were performed, i.e. for Sed 1, 2 and 3;

§ inhibitions < 0 (i.e. stimulations) were set to 0% inhibition; a L5 didn´t test D1, therefore no % inhibitions for D1 available

# Sediment 1, 5 and 7 (in italics): no invalid tests, hence same results as based on all data

| | **Bacteria Luminescence Test, % Inhibition in D1, Pore Water** | | | | | | |
|---|---|---|---|---|---|---|---|
| sample | **Sed 1** | **Sed 2** | **Sed 3** | **Sed 4** | **Sed 5** | **Sed 6** | **Sed 7** |
| number of laboratories | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| invalid measurements | 0 | 2 | 1 | 1 | 0 | 1 | 0 |
| % invalid measurements | 0,0 | 20,0 | 10,0 | 20,0 | 0,0 | 20,0 | 0,0 |
| | **Laboratory specific % inhibition in D1, including invalid data §** | | | | | | |
| L01 | 16,2 - 15,2 | 3,5 - 1,8 | 4,6 - 8,3 | 27,0 | 32,9 | 0 | 11,4 |
| L02 | 42,3 - 47,6 | 17,6 - 17,7 | 13,9 - 15,2 | 32,0 | 27,7 | 12,1 | 28,4 |
| L03 | 57,1 - 58,7 | 31,8 - 33,1 | 31,2 - 28,3 | 87,4 | 72,2 | 41,8 | 62,9 |
| L04 | 0 - 0 | 0 - 0 | 0 - 0 | 0 | 0 | 57,7 | 10,9 |
| L05 a | - | - | - | - | - | - | - |
| L06 | 31,0 - 29,9 | 27,5 - 18,0 | 23,4 - 10,3 | 4,6 | 1,3 | 0 | 18,8 |
| min / max | 0 - 58,7 | 0 - 33,1 | 0 - 31,2 | 0 - 87,4 | 0 - 72,2 | 0 - 57,7 | 10,9 - 62,9 |
| | **% inhibition in D1, results, including invalid data** | | | | | | |
| number of laboratories for statistics | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| number of measurements for statistics | 10 | 10 | 10 | 5 | 5 | 5 | 5 |
| **mean** | **29,8** | **15,1** | **13,5** | **30,2** | **26,8** | **22,3** | **26,5** |
| 95% CI | 1,3 - 58,3 | 0 - 32,3 | 0 - 28,0 | 0 - 73,4 | 0 - 63,4 | 0 - 54,8 | 0 - 53,2 |
| 95% PI | 0- 74,8 | 0 - 42,3 | 0 - 36,4 | 0 - 98,5 | 0 - 84,5 | 0 - 73,5 | 0 - 68,7 |
| NW (95%CI / mean) | 1.9 | 2.1 | 2.1 | 2.4 | 2.4 | 2.5 | 2.0 |
| sr (repeatability)* | 1,8 | 3,1 | 4,4 | n.d. | n.d. | n.d. | n.d. |
| sL (inter-laboratory variability) | 22,9 | 13,5 | 10,8 | 34,8 | 29,4 | 26,1 | 21,6 |
| sR (reproducibility) | 23,0 | 13,9 | 11,7 | 34,8 | 29,4 | 26,1 | 21,6 |
| sr % | 6,1 | 20,4 | 32,7 | n.d. | n.d. | n.d. | n.d. |
| sL % | 76,9 | 89,5 | 80,1 | 115,3 | 109,8 | 117,1 | 81,4 |
| **sR %** | **77,1** | **91,8** | **86,5** | **115,3** | **109,8** | **117,1** | **81,4** |
| | **Laboratory specific % inhibition in D1, only valid data § #** | | | | | | |
| L01 | *16,2 - 15,2* | 3,5 - 1,8 | 4,6 - 8,3 | 27,0 | *32,9* | 0 | *11,4* |
| L02 | *42,3 - 47,6* | 17,6 - 17,7 | 13,9 - 15,2 | 32,0 | *27,7* | 12,1 | *28,4* |
| L03 | *57,1 - 58,7* | 31,8 - 33,1 | 31,2 - 28,3 | 87,4 | *72,2* | 41,8 | *62,9* |
| L04 | *0 - 0* | 0 - 0 | 0 - 0 | 0 | *0* | 57,7 | *10,9* |
| L05 a | - | - | - | - | - | - | - |
| L06 | *31,0 - 29,9* | - - | - 10,3 | - | *1,3* | - | *18,8* |
| min / max | *0 - 58,7* | 0 - 33,1 | 0 - 31,2 | 0 - 87,0 | *0 - 72,2* | 0 - 57,7 | *10,9 - 62,9* |
| | **% inhibition in D1, results, only valid data #** | | | | | | |
| number of laboratories for statistics | *5* | 4 | 5 | 4 | *5* | 4 | *5* |
| number of measurements for statistics | *10* | 8 | 9 | 4 | *5* | 4 | *5* |
| **mean** | ***29,8*** | **13,2** | **12,4** | **36,6** | ***26,8*** | **27,9** | ***26,5*** |
| 95% CI | *1,3 - 58,3* | 0 - 37,1 | 0 - 27,2 | 0 , 94,9 | *0 - 63,4* | 0 - 70,1 | *0 - 53,2* |
| 95% PI | *0- 74,8* | 0 - 42,6 | 0 - 35,7 | 0 108,5 | *0 - 84,5* | 0 - 79,1 | *0 - 68,7* |
| NW (95%CI / mean) | *1.9* | 2.8 | 2.2 | 2.6 | *2.4* | 2.5 | *2.0* |
| sr (repeatability)* | *1,8* | 0,8 | 1,7 | n.d. | *n.d.* | n.d. | *n.d.* |
| sL (inter-laboratory variability) | *22,9* | 15,0 | 11,8 | 36,7 | *29,4* | 26,5 | *21,6* |
| sR (reproducibility) | *23,0* | 15,0 | 11,9 | 36,7 | *29,4* | 26,5 | *21,6* |
| sr % | *6,1* | 5,7 | 13,9 | n.d. | *n.d.* | n.d. | *n.d.* |
| sL % | *76,9* | 113,7 | 94,6 | 100,2 | *109,8* | 95,0 | *81,4* |
| **sR %** | ***77,1*** | **113,7** | **65,6** | **100,2** | ***109,8*** | **95,0** | ***81,4*** |