Experimental power of laboratory-scale results and transferability to full-scale anaerobic digestion
Maximilian Lüdtke, Åke Nordberg and Christian Baresel

ABSTRACT
Anaerobic digestion is today internationally acknowledged as an environmentally sound process for energy and nutrient recovery from organic wastes, and it is the dominant sludge treatment technology in most countries’ wastewater treatment plants. Laboratory- or pilot-scale experiments are commonly used as a first step to investigate the potential of new ideas or to confirm research hypothesis before confirmation in full-scale. The objectives of this study were to investigate transferability of methane yield assessments between laboratory- and full-scale, and to compare the influence of experimental uncertainties on experimental power in parallel continuous digester experiments for the two scales. Both batch experiment data (used in a simple equation), as well as continuous laboratory experiments, could be used to predict full-scale methane yield with a high accuracy (<5% difference). Full-scale digesters significantly outperformed hand-fed laboratory digesters in terms of experimental power regarding relative differences in methane yield between two digesters operated in parallel. However, to justify costly long-term continuous laboratory-scale experiments with sufficient experimental power and potentially high transferability, resources also have to be allocated to measures that ensure a high data quality from full-scale reference facilities.

Key words | anaerobic digestion, full-scale, laboratory-scale, sewage sludge, transferability

INTRODUCTION
Anaerobic digestion (AD) is today internationally acknowledged as an environmentally sound process for energy and nutrient recovery from organic wastes (IEA 2016). AD is the dominant sludge treatment technology in most countries’ wastewater treatment plants (WWTPs) (Bachmann et al. 2015).

Research on WWTP AD is motivated by the process’s strong influence on the WWTPs’ economic as well as environmental balance sheets. Typical research questions include, but are not limited to: (i) substrate evaluation (e.g. methane yield, nutrient content, toxicity (Angelidaki et al. 2009)), (ii) impact on other process steps in the WWTP (e.g. digestate dewatering performance, amount and composition of compounds that are released back to the WWTP for treatment (Bivins & Novak 2001)), (iii) microorganism activity (e.g. rate and ability to degrade various unwanted compounds like pharmaceuticals and organic pollutants, adaption to inhibitory substances (Institute for Environment and Sustainability 2002)), and (iv) overall process robustness to overloading incidents (De Vrieze et al. 2013).

Experiments are typically performed in three different scales: laboratory, pilot-, and full-scale, where verifiable improvements in full-scale performance are considered the ultimate proof for a successful new technology and process. However, full-scale experiments are often not justifiable for a number of reasons, for example: (i) capital costs associated with full-scale experimental construction, equipment, and instrumentation are high, as are operational costs (salaries, energy and/or materials) during the long evaluation times typically required in AD research (several months); (ii) environmental and sanitary risk: AD constitutes an integral part of WWTPs, which cannot be jeopardized by pushing the process limits; (iii) data quality: as state-of-the-art AD facilities approach theoretical limits of efficiency, any experimental effects due to new process improvements will become harder to separate from noise introduced by factors from the world outside of the experiments. Consequently, a majority of tests of new process improvements begin in laboratory-scale, where costs can be kept low and experimental
accuracy can be kept high through a high degree of control of the process variables.

A laboratory-scale batch test is the most commonly used tool for evaluation of a substrate’s methane yield (through biochemical methane potential (BMP) tests), content of inhibitory substances, and for gathering kinetic degradation data that can be used for modeling and simulations. In a typical batch test, digester material is mixed with a substrate (or several, in the case of co-digestion) in a closed vessel, and evolution of biogas and/or other interesting parameters is recorded over time. The BMP value, measured as volume methane per weight organic matter (e.g. chemical oxygen demand) or volatile solids (VS)), is considered as the maximum yield of a substrate. BMP is registered once the accumulated biogas production is considered to have plateaued (Angelidaki et al. 2009), or more formally, when the daily biogas production has dropped to less than 1% (Holliger et al. 2016) or 0.5% (Verein Deutscher Ingenieure 2016) of the accumulated biogas production in three consecutive days. However, the limitation of performing only batch type experiments is that they can only provide a snapshot of the expected full-scale behavior. To solve this, kinetics and levels of inhibition derived from batch experiments can be translated into mathematical expressions and fed into computer models, which then in the best case can provide a reasonable dynamic prediction of the full-scale behavior. This approach has proven to be successful even at plant-wide modeling (Batstone et al. 2015; Arnell et al. 2016) with a majority of implementations being based on the International Water Association’s Anaerobic Digestion Model 1 (ADM1) (Batstone et al. 2002).

Continuous laboratory- or pilot-scale experiments are applied when batch tests are not sufficient to answer the research question. This relates to topics such as adaptation of AD to novel substrates and inhibitory compounds over time, evaluation of transient behavior due to larger process changes (e.g. change of operational temperature, and transition from single- to multi-stage digestion), influence of mixing and feeding frequency, accumulation/depletion of certain micro-nutrients, and impact of solids retention time manipulation. The most common setup for such experiments is where results from digesters operating in parallel are compared with each other, as this allows quantification of improvements to be formulated as an experimental effect that a promising new treatment/process might have on a parameter of interest when applied to a test system compared with an untreated control system: ‘Process A is X % better than Process B’. In most cases, the experimental effect is either directly or indirectly evaluable through measured differences in specific methane yield (SMY [NL kgVS⁻¹]) and/or degree of digestion (DD, i.e. VS conversion [%], further described in the Methods section).

Measurement and sampling errors caused by e.g. instrumentation, human error, and substrate heterogeneity introduce noise in both batch and continuous experiments (Esbensen & Paasch-Mortensen 2016). Such noise limits the detectable experimental effect and thereby lowers the experimental power (see Methods section for the definition of experimental power that is used in this text). Even though experimental noise in many cases can be compensated for by increasing the number of replicates and/or the total evaluation time, there will always be a trade-off where resources spent on experiments have to be weighed against not only the expected experimental power, but also the results’ transferability to full-scale. This also influences the choice of experimental setup, as a relatively high accuracy obtained in laboratory-scale can be meaningless if the transferability to full-scale is low.

There is a large variation in reported transferability of SMY in peer-reviewed literature, e.g.: SMY was overestimated by 19–24% in 6 L semi-pilot digesters when compared with 300 L digesters treating industrial and food waste (Ruffino et al. 2015; Fiore et al. 2016). Koch et al. reported a 3% higher SMY when batch tests were compared with long-term (one year) full-scale digestion of a primary and secondary sludge mixture (Koch et al. 2016). A comparison of degradability of thermally hydrolysed sludge did not reveal significant differences between laboratory-scale batch digestion and full-scale operation (batch digestion tended to underestimate degradability by approximately 10%) (Batstone et al. 2009).

The objective of this article was to identify limitations and possibilities in using data from laboratory-scale experiments (batch and semi-continuous) and data from WWTP full-scale AD operation to quantify experimental effects. The evaluation was focused on two topics: (i) transferability of methane yield assessments between BMP tests, semi-continuously operated laboratory-scale digesters and full-scale AD; (ii) comparison of uncertainties, and their influence on the experimental power in the determination of relative differences in methane yield between two digesters operated in parallel in laboratory- or full-scale.

**METHODS**

**Datasets**

For the evaluations, a total of five datasets from two of Stockholm’s largest WWTPs, Käppalaverket and
Himmerfjärdsverket, Sweden, were used (overview of the datasets is presented in Table 1). Four datasets were obtained from Käppala WWTP using primary sludge from the WWTP as the substrate: Käppala full-scale year 1 (K1), Käppala full-scale year 2 (K2), Käppala batch BMP (Batch), Käppala laboratory-scale (semi-continuously hand fed digesters) (KL). K1, KL, and Batch datasets were obtained during a 23-week period when KL digesters were operated in parallel with K1. K2 features full-scale data from the same digester as in K1, acquired in the same manner and under similar operational conditions one year later and after a gas meter change. The fifth dataset was obtained from full-scale operation of three parallel digesters at Himmerfjärdsverket WWTP (HF). All data from the continuous experiments (K1, K2, KL, and HF) came from weeks with stable operation defined according to US EPA guidelines for sludge digestion as a low volatile fatty acids to alkalinity ratio (<0.35) and a stable gas composition (>60% CH4) (US EPA 1976). Once a week, the digester total solids (TS) and VS were measured according to Swedish Standard (Swedish Institute for Standards 2000). Detailed information about the individual datasets follows below.

**Batch**

Batch data were obtained from a 26-day BMP test that fulfilled all the compulsory elements for validation according to Holliger et al. (2016). Methane production normalized to STP (101.325 kPa, 273.15 K) was registered using AMPTSII (Bioprocess Control, Sweden, see manufacturer home page for a picture of the setup) after the gas had been passed over a 3 M NaOH solution to capture CO2. Substrate:inoculum VS-ratio was 1:2 by weight. Time from the harvest of inoculum (digester content from the full-scale process, TS: 2.55% and VS: 70.5%) to experiment start was less than three hours. Substrates were cooled (24 h, 4 °C) and thawed frozen (24 h, −18 °C) primary sludge, taken from the WWTP (TS: 6.17%, VS: 88.2%). Frozen sludge was included as a control for KL experiments where fresh sludge could not be used due to practical reasons. Control triplicates included inoculum only (i.e. ‘blank’) and microcrystalline cellulose (SERVA Electrophoresis GmbH, Germany) as a positive control. Active volume was approximately 300–400 mL in 600 mL glass bottles. Methane production from the fed triplicates was reported after the inoculum’s production had been subtracted. N2 was used for batch bottle headspace flushing. The overestimation of registered methane production due to the flush gas was compensated for by the AMPTSII software. The potential risk of

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Overview of the datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dataset (Digester names)</strong></td>
<td>Scale</td>
</tr>
<tr>
<td>Batch</td>
<td>Laboratory, Batch</td>
</tr>
<tr>
<td>KL (KL1, KL2)</td>
<td>Laboratory, Semi-continuously hand fed</td>
</tr>
<tr>
<td>K1</td>
<td>Full-scale</td>
</tr>
<tr>
<td>K2</td>
<td>Full-scale</td>
</tr>
<tr>
<td>H (HD1, HD2, HD3)</td>
<td>Full-scale</td>
</tr>
</tbody>
</table>

HRT, hydraulic retention time; OLR, organic loading rate.
BMP underestimation (averaging approximately 14% (Koch et al. 2015)) due to the use of pure N₂ as opposed to a CO₂-containing flush gas was considered to be negligible for all the reported BMPs, as even the inoculum production alone was considerably higher in this study (100 NL kgVS⁻¹) than the BMPs reported by Koch et al. (2015) (<50 NL kgVS⁻¹).

**Semi-continuously fed laboratory-scale digesters**

KL data were obtained from a 23-week period with two 1.8 L active volume (Vactive) mechanically stirred laboratory digesters operated in parallel (Bioreactor Simulator, Bioprocess Control, Sweden, see manufacturer homepage for a picture of the setup). Feeding was performed manually once each workday using pre-weighed thawed portions of frozen primary sludge (mean TS: 5.5%, and mean VS: 86.8%). A weekly OLR of 3 kgVS m⁻³ d⁻¹ was maintained through distributing the total seven days’ feeding as follows: Mondays through Thursdays – one daily portion plus Sunday’s feed divided equally over the four days; Fridays – two daily portions (i.e. including Saturday’s feeding). Median hydraulic retention time (HRT) was 16 days. Methane production was registered as described for Batch experiments.

**Full-scale digesters**

K1 and K2 data were obtained from two 23-week time periods in two consecutive years and feature normal operation of a full-scale digester treating primary sludge only (a second digester of equal size at the WWPT treats the effluent of the first digester and all secondary sludge, but this digester was not included in this study). During acquisition of gas data for K1, a gas flow meter with mass-flow/thermal-conduction measurement principle was used. The gas flow meter was changed to an ultrasonic meter (XGM868i, General Electrics) before data acquisition for K2 had begun. The change was performed in order to mitigate suspected problems associated with the thermal mass flow measurement principle according to the authors’ previous experience of similar installations (condensation and fouling affecting the heating element and the temperature sensor, caused by the raw biogas). Median of key operational parameters for K1 and K2, respectively: Vactive, 7,500 and 7,600 m³; HRT, the median 18 days HRT. Feeding of the HF digesters was performed in a way that allowed application of an equal median OLR of 2.7 kgVS m⁻³ d⁻¹ to all three digesters: three substrate streams (external substrate, primary, and secondary sludge) were pumped into a common feeding pipe. This feeding pipe was connected to each digester via a separate time controlled valve that was allowed open for one reactor at a time during a one-hour window. Gas flow was registered using ultrasonic meters (B200, Endress + Hauser, Switzerland).

**Methane yield determination**

Experimental data from Batch, KL, K1, and K2 were used to compute and compare different methods of determining SMY. Difference in SMY between cooled and frozen sludge triplicates in Batch was not significant at α = 0.05 in a Wilcoxon rank sum test (medians 391 and 379 NL kgVS⁻¹, respectively, p > 0.2). Therefore, the reported batch data in this article reflect the pooled results from these two triplicates. Sludge flows and digester Vactive were taken directly from the online Supervisory Control and Data Acquisition (SCADA) system for K1 and K2 but measured and computed by hand (weighing) for KL and Batch, assuming a sludge density of 1 kg dm⁻³. As an additional method of calculating full-scale methane yield, the following formula was used:

\[ \frac{\text{BMP}(t)}{\text{HRT}} e^{-1/\text{HRT}} dt \]

where BMP(t) represents the median accumulated methane yield from Batch up to day t (BMP (t > 26) = BMP(t = 26)), which is multiplied with the fraction of substrate that had a residence time t, taken from the residence time distribution function of an ideal continuously stirred tank reactor with K1’s median HRT, according to Levenspiel (2012). The digester was considered to be well-mixed with a total mixing energy of 7 W m⁻³ active volume (combination of mechanical stirring and circulation pumps) according to US EPA (1987) guidelines.

Calculations of DD for KL, K1, and K2 were performed according to the Van Kleeck method (Van Kleeck 1945).

**Comparison of experimental power in parallel digester experiments**

In this text, the whole experimental setup was considered and ‘experimental power’ was interpreted as ‘the smallest
statistically significant detectable difference in SMY between two digesters operating in parallel. Impact of variability within digester pairs operating in parallel on experimental power was estimated numerically based on experimental data from KL and H. This was done in order to visualize differences in experimental resolution between full-scale and laboratory-scale.

Comparisons were made between all possible digester pairs within each dataset (i.e., for KL: KL1 and KL2, and for H: HD1 and HD2, HD1 and HD3, HD2 and HD3). The procedure for each digester pair comparison, digester 1 (D1) and digester 2 (D2), is described below:

1. A normalized dataset was created where any systematic difference between digesters D1 and D2 was filtered out by multiplying all 23 weekly median digesters SMYs (WM-SMYs) in the experimental dataset for D1 with a constant in order to reach median (D1 WM-SMYs) = median (D2 WM-SMYs) for that 23-week period.

2. ‘Test periods’ were created by randomly extracting (with replacement) 7 weeks from the normalized 23 weeks dataset. A total of 100 7-week test periods were created. Seven weeks (=49 days) was chosen as it approximates a typical experimental evaluation time of 3 HRTs of stable operation for a commonly used 16-day digestion process (=48 days).

3. The differences in median (WM-SMYs) between D1 and D2 for each 7-week test period were saved to a list, ‘Start difference’.

4. For each of the 100 test periods, repeated tests for significant difference in median (WM-SMYs) between D1 and D2 were performed iteratively with an increasing 1.0X multiplication factor applied to the seven WM-SMYs of D1 (i.e. to simulate an X% yield-increasing experimental effect affecting only D1). The multiplication factor was increased until \( p < 0.05 \) was reached in a two-sided pairwise Wilcoxon rank sum test with H0: median (D1 WM-SMYs) = median (D2 WM-SMYs), and H1: median (D1 WM-SMYs) \neq \) median (D2 WM-SMYs). The differences in median WM-SMYs between D1 and D2 for each test period after significant difference had been reached were saved to a list, ‘End difference’.

5. Start difference and End difference results for each of the 100 7-week test periods were compared. The results were expressed as the percentage of applied simulated experimental effect that was required to reach statistical difference \( (p < 0.05) \) in median (WM-SMYs) when D1 was compared with D2.

RESULTS AND DISCUSSION

Transferability and uncertainties of methane yield determination

The investigated datasets K1 and K2 from Käppala WWTP show, not surprisingly, that full-scale AD data can comprise large relative variations in sludge inflow (Figure 1(a)), sludge VS (Figure 1(b)) and methane flow (Figure 1(c)) when examined on shorter time scales such as weeks.

Seen over a longer time, however, K1 and K2 median inflow and VS content were similar (K1: 24.8 m³ h⁻¹, 95% confidence interval (CI) ± 2.1% and 87%, CI ± 1%, K2: 28.5 m³ h⁻¹, CI ± 4% and 88%, CI ± 1%), which indicates a sound basis for comparisons between these two time periods with respect to OLR and HRT. Median registered methane flow was, on the other hand, larger in K2 than in K1 (K1: 368 Nm³ h⁻¹ ± 2.3%, K2: 418 Nm³ h⁻¹, CI ± 1.5%), most likely due to the change of gas flow meters.

An overview of experimentally determined SMYs and DDs of the datasets Batch, KL, K1, K2, and calculated full-scale SMY from Equation (1) is presented in Figure 2.

Our data showed an expected higher median SMY from Batch (i.e. BMP) (382 NL kgVS⁻¹, CI ± 2%) compared with the other datasets. SMY from KL (341 NL kgVS⁻¹, CI ± 3%) and SMY calculated with Equation (1) (355 NL kgVS⁻¹) did not differ significantly from K2 (551 NL kgVS⁻¹, CI ± 6%), while K1 SMY (285 NL kgVS⁻¹, CI ± 7%) was significantly lower. Uncertainties in assessments of median SMY were thereby the lowest in Batch followed by KL and were the highest in full-scale datasets K1 and K2. This is accordance with previously observed difficulties in estimating SMY from full-scale data (Batstone et al. 2009).

DD was higher in the full-scale datasets (K1: 65%, K2: 67%) compared with KL (63%), which indicates that SMYs derived from semi-continuously hand-fed digester experiments are at risk of resulting in an underestimation of full-scale SMY. This is further strengthened by the fact that KL data were obtained at both a lower OLR and a higher HRT than K1 and K2; this difference in OLR and HRT between the two scales would, if anything, have been expected to result in a higher SMY for KL. Uncertainties in the estimated DD were similar in the laboratory- and full-scale, which can be expected, as the major contributing error sources behind this parameter were the same (i.e. manual sampling and analysis of TS and VS). DD was also used together with SMY in order to justify the accuracy of the new gas flow meter that was in use during the time
period of K2 compared with the old meter that was operational in K1: by forming the ratio median (SMY)/median (DD) for KL, K1, and K2, the laboratory-scale results in KL could be related to one another under the reasonable assumption that a similar gas production per kg VS degraded should be expected from a certain VS conversion. KL’s and K2’s SMY/DD ratios were nearly identical (KL: 546 NL kg\(^{-1}\) VS degraded and K2: 527 NL kg\(^{-1}\) VS degraded), while SMY/DD was considerably lower for K1 (438 NL kg\(^{-1}\) VS degraded\(^{-1}\)). Based on this close agreement, we argue

Figure 1 | (a) Sludge inflow, (b) VS content (% of wet weight), and (c) methane gas flow, from datasets K1 and K2, obtained from the same time period in consecutive years.

Figure 2 | (a) Comparison of SMY from datasets Batch, KL, K1, K2. Crossing horizontal line represents SMY calculated from batch kinetics assuming ideally mixed full-scale digester. (b) Comparison of DD from datasets KL, K1, K2. Points represent weekly medians. Boxplots: Whiskers – highest/lowest observation within 1.5 interquartile range; Hinges – 25th and 75th quartiles; Thick line – median; Notches span a 95% CI of the median. Note: Truncated y-axis.
that the SMY calculated from K2 data using the new ultrasonic gas flow meter better reflected the true SMY for the full-scale operation.

In summary, our results indicate that both data from continuous laboratory-scale experiments, as well as a simple batch BMP experiment together with a simple calculation (Equation (1)), can give a very good approximation of the expected full-scale SMY yield (<5% difference in the studied datasets), assuming that the full-scale digester is properly mixed. In this sense, methane potential tests can be seen as more resource efficient compared with the workload, time and cost invested in continuous experiments if only the prediction of methane yield is of interest. However, if the new substrate can be assumed to require an adaption of the microbiology, and/or dynamic or long-term process response is of interest, a more time-consuming and resource-intensive laboratory-scale continuous digestion experiment should be considered.

**Comparison of power in parallel digester experiments**

Operating replicate digesters in parallel is the best approach if the experimental aim is to quantify relative differences in performance between a test digester and a control digester. A comparison of replicate digester performance in full- and laboratory-scale is shown in Figure 3, which features data from three full-scale reactors in the H dataset and the two hand-fed laboratory-scale reactors in KL. Note that full-scale and laboratory-scale data in this figure come from different experiments, and yield in absolute numbers can therefore not be compared between laboratory- and full-scale, but only inter- replicate differences within each dataset.

From examining Figure 3(a), it is clear that the weekly evolution of SMY within the replicates differs between laboratory- and full-scale, i.e. an increase or decrease in weekly yield in one replicate co-varies more strongly with the other replicate(s) in full-scale. Pearson’s pairwise correlation was weak between the laboratory-scale replicates \( r = 0.34 \), and high between the full-scale replicates \( r > 0.96 \). It is evident that the error sources associated with the hand-feeding of the laboratory-scale reactors (sampling error due to smaller volumes, manual measurements instead of automated online instrumentation, etc.) introduced an increased uncertainty seen over the short term. However, these uncertainties eventually evened out as the total experiment time increased (difference in median weekly SMY < 2.3%), which indicates that minimal systematic bias had been introduced due to the manual feeding procedure. The impact of such uncertainties on the experimental power of the laboratory-scale for a typical 3-HRT evaluation period is visualized in Figure 3(b). In this figure, the median power of the full-scale is improved close to three-fold compared with the laboratory-scale.

![Figure 3](https://iwaponline.com/wst/article-pdf/76/4/983/451207/wst076040983.pdf)
Interestingly, this is opposite to the trend displayed in Figure 2, where the uncertainties in the assessment of absolute SMY were larger in the full-scale compared with the manually fed laboratory-scale, and highlights the strengths of the automated parallel continuous experiment design with a common feeding pipe that was used in H when differences in relative SMY are of interest.

For continuous parallel digester experiments, our results indicate that manual feeding of digesters with a moderately heterogeneous substrate such as primary sludge requires preferably both experienced personnel and evaluation times longer than the commonly used 3 HRT in order to reach sufficient experimental power to detect experimental effects <10% when they are measured as relative differences in SMY.

The transferability of laboratory-scale results to full-scale operation is dependent on reliable data from full-scale facilities, indicated by e.g. the improved transferability of laboratory-scale SMY when changing gas meter between K1 and K2. Thus, before costly long-term continuous laboratory-scale experiments with sufficient experimental power and potentially high transferability can be justified, resources also have to be allocated to measures that ensure high-quality data from full-scale digester(s).

CONCLUSIONS

The investigated data from Käppala WWTP have shown that a simple method of estimating full-scale methane yield utilizing BMP data, as well as bench-scale digester experiments, could be used to predict expected full-scale SMY with a high accuracy (<5% difference). Furthermore, our results indicate that the increased control of operational variables (e.g. HRT and OLR) offered in semi-continuously hand-fed laboratory-scale experiments can lower uncertainties in assessments of absolute SMY substantially compared with a single full-scale digester. This directly translates to shorter evaluation times and thereby lower experimental costs.

However, the results from our comparison of parallel digester operation indicate that automated feeding in full-scale can significantly outperform hand-fed laboratory-scale in terms of experimental power in the detection of relative differences in methane yield between a test and a control digester.

Moreover, with regard to general transferability, this study highlights the importance of resource allocation to improvement of full-scale data quality before costly laboratory experiments can be justified.

FUNDING

This study was co-financed by IVL Swedish Environmental Research Institute together with Käppala Association, and Syvab.

REFERENCES


Holliger, C., Alves, M., Andrade, D., Angelidaki, I., Astals, S., Baier, U., Bougrier, C., Buffiere, P., Carballa, M., de Wilde, V.,


