Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020

Xavier Bouthillier, Gaël Varoquaux

Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020

Xavier Bouthillier   XAVIER.BOUTHILLIER@UMONTREAL.CA
Mila
Montréal, QC, Canada

Gaël Varoquaux   GAEL.VAROQUAUX@INRIA.FR
Inria Saclay
Palaiseau, France
and
Mila
Montréal, QC, Canada

Abstract

How do machine-learning researchers run their empirical validation? In the context of a push for improved reproducibility and benchmarking, this question is important to develop new tools for model comparison. This document summarizes a simple survey about experimental procedures, sent to authors of published papers at two leading conferences, NeurIPS 2019 and ICLR 2020. It gives a simple picture of how hyper-parameters are set, how many baselines and datasets are included, or how seeds are used.

1. Introduction

Experiments play an increasing role in machine learning research. The prototypical experimental paradigm is to measure the generalization error of a model on a benchmark dataset, and, often, to compare it to baseline models. Yet, trustworthy experimentation is difficult: thorough benchmarking has shown that classic baselines outperform later work initially reported as improvement for knowledge base completion (Kadlec et al., 2017), neural language models (Melis et al., 2018), or generative adversarial networks (Lucic et al., 2018). Empirical results are affected by arbitrary factors such as the random seeds of neural-network fitting procedures (Bouthillier et al., 2019) or non-deterministic benchmarking environments in reinforcement learning (Henderson et al., 2018). Computational budget matters for a fair comparison (Lucic et al., 2018; Melis et al., 2018), in particular for selection of model hyperparameter (Dodge et al., 2019).

The growing recognition of these challenges in machine learning research has motivated research in better procedures for experimentation and benchmarking: better hyperparameter tuning (Bergstra and Bengio, 2012; Hansen, 2006; Hutter et al., 2018; Li et al., 2018; Snoek et al., 2012), controlled statistical testing of model performance (Bouckaert and Frank, 2004; Demšar, 2006; Dietterich, 1998; Hothorn et al., 2005). To reach best the community, ideally these developments should fit as well as possible the current practices.

Here, we present a survey of experimental procedures currently used by practitioners at two of the leading conferences, NeurIPS2019 and ICLR2020. The survey was conducted by

asking simple anonymous questions to the corresponding authors of the papers published at these venues. We detail the results of the survey and provide a light analysis[1].

Highlights A vast majority of empirical works optimize model hyper-parameters, thought almost half of these use manual tuning and most of the automatic hyper-parameter optimization is done with grid search. The typical number of hyper-parameter set is in interval 3-5, and less than 50 model fits are used to explore the search space. In addition, most works also optimized their baselines (typically, around 4 baselines). Finally, studies typically reported 4 results per model per task to provide a measure of variance.

2. Survey methodology

The objective of the survey is to provide a portrait of the current common practices for experimental design in the machine learning community. We selected papers at the peer-reviewed conferences NeurIPS 2019 and ICLR 2020 as a representative sample of practices considered valid by the reviewers.

We modeled our questions to capture benchmarking methods, the predominant experimental procedure to measure the performance of a new algorithm. The survey is limited to 10 short and simple questions, all multiple choices. This was important to favor high response rates from the researchers, and should provide valuable information nevertheless considering the lack of documentation on the topic. The survey was also anonymous to favor high response rates.

First question serves as a filter of empirical and theoretical papers, second one as a filter for questions on hyperparameter optimization. The third, fourth and fifth questions measure the popularity of methods as well as the search space in terms of dimensionality and exploration. Sixth question measures comparability of models in benchmarks, while seventh and eight ones measure the number of comparisons and variety of benchmarks. Finally ninth question quantifies sample size upon which conclusions are made and tenth question measures reproducibility of the papers with sample size 1.

For NeurIPS, all author names, paper title and PDF were collected from the pre-proceeding webpage[2]. The emails were then automatically collected from the PDFs, with manual intervention when required. For ICLR, all author names, paper title and emails were collected using the official OpenReview library[3].

The survey for NeurIPS was sent on Thursday, 12th December, during the conference. A second email to remind authors about the survey was sent on Wednesday, 18th December. The response rate promptly raised from 30% to 34% after the reminder. The survey for ICLR was sent on Monday, 23th December, 4 days after the paper decision notifications. Considering the high response rate, no second email was sent to remind authors of ICLR.

3. Results

The questionnaire was sent to all first authors or corresponding authors of the accepted papers at NeurIPS2019 and ICLR2020. The response rates were 35.6% $\left(\frac{452}{1269}\right)$ and 48.6%

1. Anonymous data is available at github.com/bouthilx/ml-survey-2020
2. https://papers.nips.cc/book/advances-in-neural-information-processing-systems-32-2019
3. https://github.com/openreview/openreview-py

$\left(\frac{334}{687}\right)$ for NeurIPS2019 and ICLR2020 respectively. On each figure below, the total number of answers is given at the top, while the number of replies for a given answer is given in the bars. The confidence intervals for all results are below 5% for a confidence level of 95%.
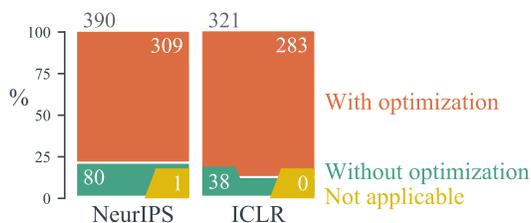
Question 1)
Did you have any experiments in your paper? If no, you are already done with the survey, thank you and have a good day. :)
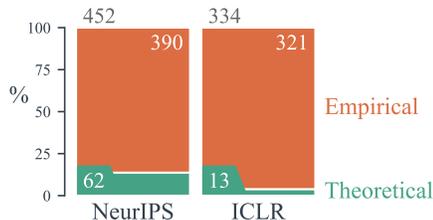
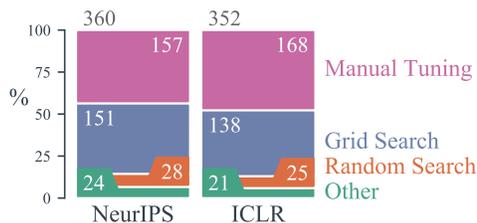**Question 2)**
Did you optimize your hyperparameters?

Results are for empirical papers only.
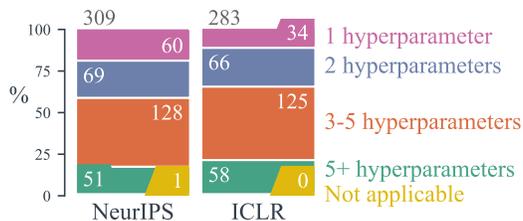
Question 3)
If yes, how did you tune them?

Results are for empirical papers with optimization only. Papers may use more than one method, hence it sums to more than the number of empirical papers.

Question 4)
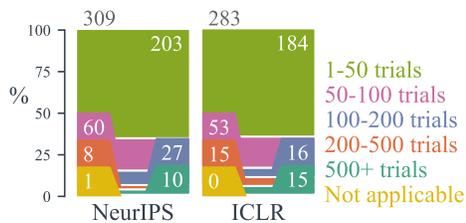How many hyperparameters did you optimize?

Results are for empirical papers with optimization only.

Question 5)
How many trials/experiments in total during the optimization? (How many different set of hyperparameters were evaluated)
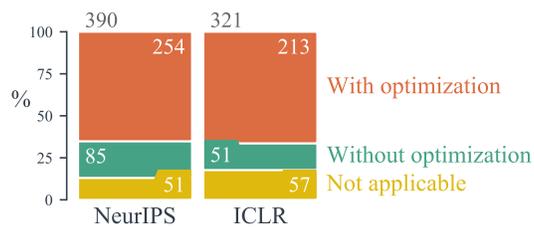
Results are for empirical papers with optimization only.

Question 6)
Did you optimize the hyperparameters of your baselines? (The other models or algorithms you compared with)
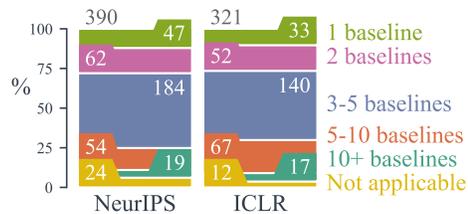
Results are for empirical papers only.

Question 7)
How many baselines (models, algos) did you compare with? (If different across datasets, please report maximum number, not total)

Results are for empirical papers only.

Question 8)
How many datasets or tasks did you compare on?

Results are for empirical papers only.
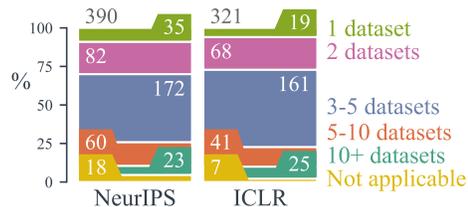
Question 9)
How many results did you report for each model (ex: for different seeds)

Results are for empirical papers only.
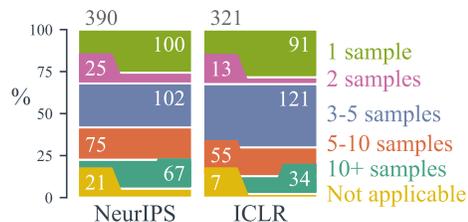
Question 10)
If you answered 1 to previous question, did you use the same seed in all your experiments? (If you did not seed your experiments, you should answer 'no' to this question.)

Results are for empirical papers with 1 result per model only.
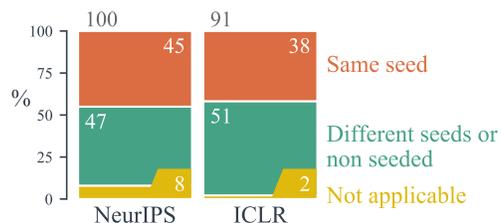
4. Discussion

4.1 Analysis

We first note the results are similar for NeurIPS and ICLR across all questions. One of the few significant differences being the larger proportion of theoretical papers at NeurIPS with approximately 10% more.

Hyper-parameter tuning   The most popular tuning approach is manually, but followed closely by grid search (a 1.6% difference between the two approaches at NeurIPS and 8.5% at ICLR). Together they account for more than 85% hyperparameter procedures in both conferences. The number of hyperparameters is mostly in the interval 3 to 5. The proportions are preserved if we look specifically at the answers of papers using grid search.

Figure 1: Results of question 4), **number of hyperparameters optimized**. All answers on the left, grid search only on the right.

In more than 50% of the papers, the hyperparameter search was executed on less than 50 trials. Again, the proportions are preserved if we look specifically at the answers of papers using grid search.
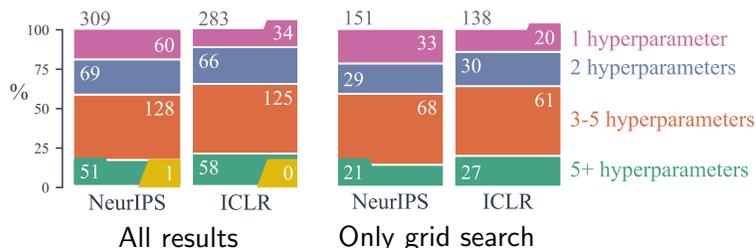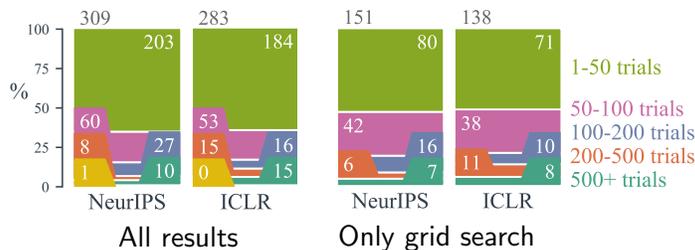
Figure 2: Results of question 5), **number of trials in total during hyper-parameter optimization**. All answers on the left, grid search only on the right.

We now focus on grid search results because it is possible to infer the number of different values for each hyperparameters which were evaluated in the papers. Using the lower bound of the number of hyperparameters (n_h) and upper bound of number of trials (n_t) reported in the survey, we can estimate an upper bound on the number of different values (n_v) evaluated in the corresponding papers with $\exp\left(\frac{\log n_t}{n_h}\right)$, given that using grid search produces the number of trials $n_t = n_v^{n_h}$. We illustrate these results in Figure 3. For more than 50% of the papers in which grid search was used, 6 or less different values were evaluated. It is important to understand that this is an upper bound, computed using $n_h = 1, 2, 3, 6$ for answers 1, 2, 3–5 and 5+ to question 4) and $n_t = 50, 100, 200, 500, 1000$ for answers to question 4). The true average number of values used must therefore be lower. As an example, changing only n_h from 3 to 4 for answer 3–5 reduces the red bar of 50% to 4 values or less.

This low number of points reveals the importance of the selection of the grid. As shown by Bergstra and Bengio (2012), grid search's performance is ensured only if the region of optimal hyperparameters is large with respect to the grid's granularity. As such, in order to ensure performance of a coarse grid search of 6 values, it would require either that the hyperparameters are optimal on a wide region, or that the prior knowledge of experimenter made it possible to design a coarse grid precisely positioned on region of optimality. In either case, hyperparameter optimization could be considered rather pointless. Considering the

5

accumulating evidence suggesting the opposite (Lucic et al., 2018; Melis et al., 2018; Dodge et al., 2019), we instead speculate that the present statistics suggest under-performing hyperparameter optimization procedures. This represents more than 20% of the empirical papers for NeurIPS and 19% for ICLR.
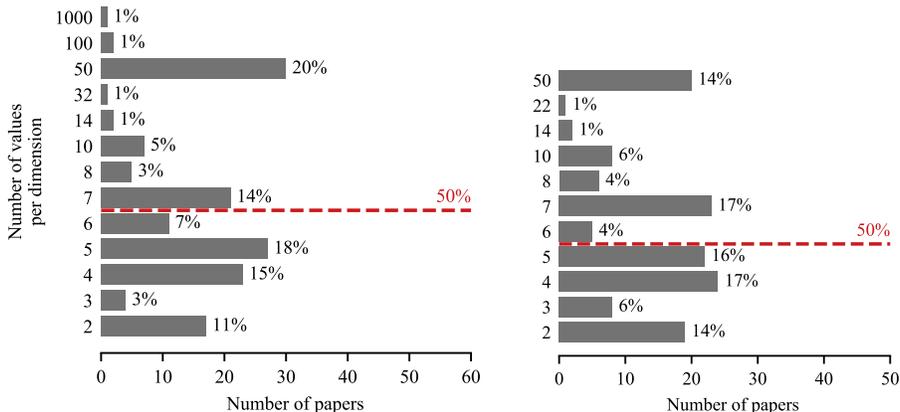

Figure 3: Number of different values for each hyperparameter

One shortcoming of the survey is the lack of information on the type of model used. Consequently, we cannot know if these numbers applies similarly to models that are computationally cheap or expensive. We would expect however that large neural networks are still more popular at ICLR than NeurIPS based on the history of these conferences. We thus believe the relative similarity of the results between the conferences to be an indication that the proportions reported here would apply to both papers on cheap and expensive algorithms.

Baselines   The majority of authors of empirical papers reported optimizing their baselines' hyperparameters. This strikes us as against our personal knowledge of recent machine learning literature. Our hypothesis is that many authors answered 'Yes' for reporting results of other papers in which the hyperparameters were optimized. This would be technically true and the question should have been formulated differently to avoid this possible confusion. The goal of this question was to measure to which extend baselines are comparable, that is, using the same experiment design as the contributed algorithm. Both the number of baselines and datasets most importantly falls in the 3 to 5 interval. We expected these numbers to be lower and suspect that it may a be bias caused by the reviewing process of the conferences we selected.

Statistical power   The most common number of reported results per model are between 1 and 5 inclusively, 26% and 37% of which falls in the interval (3, 5) for NeurIPS and ICLR respectively while 26% and 25% reported 1 result. About half of the papers with only one result per model were seeded. Statistical tests are rarely used in recent machine learning literature, nevertheless we believe it is insightful to consider what statistical power or what size effect could be detected with the sample sizes reported in this survey.

Using Neyman-Person approach for statistical testing and controlling for both error type I and error type II at 0.05, we can compute the minimal difference of performance that can be detected between two algorithms. For simplicity, lets consider a classification task in which a model have 90% accuracy on a test set of 10k samples. We can estimate the variance of the performance measure due the sampling of the test set by assuming the measure follows a binomial distribution[4]. For a sample size of 1, the minimal difference is about 1% accuracy, while it is 0.45% for sample size 5, and 0.31% for sample size 10. As 50% of sample sizes reported in the survey are below 6, the community should be concerned of benchmarks where many algorithms compared have similar performances, and make sure a sufficient sample size is used.

Opening questions on hyperparameter optimization Most studies use grid search or random search for hyperparameter optimization. This lack of diversity raises the question of why more sophisticated hyperparameter optimization methods are not used. Are researchers lacking trust in these more sophisticated methods? Are they lacking computational power and therefore favor manual procedures? Are the available codebases considered to complicated to use?

4.2 Shortcomings of the survey

Sampling bias All samples have an inherent bias. For this survey, the bias comes both from the population targeted –papers selected at two of the leading conferences, NeurIPS2019 and ICLR2020–, has well as self-selection bias in the responses: respectively 35.6% and 48.6% of researcher surveyed replied. This bias may result in sampling more researchers sensitive to the question of experimental design. As such, the bias could be considered towards better practices, implying that actual practices could be slightly less systematic then our results imply. The survey was anonymous which made it impossible to limit number of answers per paper. Consequently, unmeasured duplicate answers could also have contributed to a bias. Comparison of our results with the statistics of the reproducibility checklist at NeurIPS could serve as an estimate of these biases.

Choice of questions Results of the survey are aggregated over all empirical papers. The implications of methodologies vary significantly however for different type of tasks or different kind of models. We failed to include a question which would allow such dichotomy in the analysis. We will discuss this further in next section.

As we were collecting the answers of the survey, researchers reported some confusion on several questions.

Question 2) Many papers were not using benchmarks but rather ablation studies as their experimental paradigm. We intended questions 2-5 as measures of the exploration of the hyperparameter space and its effect beyond simple optimization. Nevertheless, some authors reported answering *not applicable* when using ablation studies. An optional textual answer for *not applicable* would have been very informative.

Question 4) There was an overlap between the intervals (1, 50), (50, 100), (100, 200) and (200-500).

---

4. Which is approximately valid unless performance is very close to 100%

Question 6) As reported by one of the respondent, the question did not ask specifically about using the same hyperparameter optimization procedure. This likely confused many authors leading them to answer *yes* if their baselines where optimized results from other papers.

Question 7) Some authors told us they answered *not applicable* because they were the first to tackle a new problem. There is generally a straw man solution that can be used as a baseline in such situation and failing to compare with it should have been considered as 0 baseline instead of *not applicable*. There was however no option '0' in the multiple choices.

Question 8) There was an overlap between intervals (3, 5) and (5, 10).

Question 9) There was an overlap between intervals (3, 5) and (5, 10).

Question 9-10) Based on the feedback received and textual answers, questions 9) and 10) were often mis-understood as seeding of model initialization in particular. The question was rather about seeding of any possible source of variation in the experiments. As a result, several authors reported to us answering *not applicable* because model initialisation was not a source of variation in their experiments. Additionally, the question did not ask specifically *per model per task*. This could have led some authors to answer number of results reported over all tasks, therefore introducing a bias upwards.

Despite these limitations, we believe current trends can be identified as we highlighted on the experimentation methodology of researchers in machine learning. Some differences are indeed large enough to be considered significant.

5. Conclusions

For reproducibility and AutoML, there is active research in benchmarking and hyperparameter procedures in machine learning. We hope that the survey results presented here can help inform this research. As this document is merely a research report, we purposely limited interpretation of the results and drawing recommendations. However, trends that stand out to our eyes are, *1)* the simplicity of hyper-parameter tuning strategies (mostly manual search and grid search), *2)* the small number of model fits explored during this tuning (often 50 or less), which biases the results and *3)* the small number of performances reported, which limits statistical power. These practices are most likely due to the high computational cost of fitting modern machine-learning models.

Acknowledgments We are deeply grateful to the participants of the survey who took time to answer the questions.

References

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–12. Springer, 2004.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734, 2019.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, 2019.

Nikolaus Hansen. The cma evolution strategy: a comparing review. In *Towards a new evolutionary computation*, pages 75–102. Springer, 2006.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3): 675–699, 2005.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Springer, 2018. In press, available at http://automl.org/book.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 69–74, 2017.

Lisha Li, Kevin G Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:1–52, 2018.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in neural information processing systems*, pages 700–709, 2018.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. *ICLR*, 2018.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Appendix A. Categorization of answers for question 10

Some authors gave textual answers to question 10. We received more textual answers than the ones reported here, but we only considered those of empirical papers for which authors answered '1' to question 9). Hence, we ignored the others. There is only 7 textual answers out of the 100 and 91 answers for question 10). The goal of question 10) was to measure to which extend papers with single results per model were reproducible. If an answer does not fit the typical *Same seed* but satisfies reproducibility, we categorize it as *Same seed*. Otherwise we categorize it as *Different seeds or non seeded*. The rationale for each decisions are described below.

> *I reported average score for each model, averaged over 10 runs. Each of the 10 runs used the same seed across all models.*

Category: Same seed

Rationale: The average estimation was seeded and thus reproducible.

> *Same seed, but estimated the variance in previous works to be small.*

Category: Same seed

Rationale: Used the same seed.

> *not applicable. there's no seed because we use pre-trained models*

Category: Same seed

Rationale: Reusing the same pre-trained model has the same effect as using the same seed to train a model.

> *Seeded the data examples, did not seed the simulations (where however enough Monte Carlo replicates were conducted to make noise negligible)*

Category: Same seed

Rationale: Assuming the noise due to simulation was indeed negligible enough, we assume seeding the sampling of data samples is enough to make the experiments reproducible.

Our paper did not involve training deep networks, and there were not any mysterious hyperparameters....

Category: Same seed

Rationale: Because the algorithm was apparently fully deterministic in a deterministic environment.

Checked that results were consistent with different seeds but no proper study of variance

Category: Different seeds or non seeded

Rationale: Because different seeds were evaluated but results for only one of them was reported.

Choice of seed has no effect on training outcome

Category: Same seed

Rationale: Assuming the authors are right and seeding has negligible effect in their experiments.