Supplementary Material
Learning to Learn: How to Continuously Teach Humans and Machines

Parantak Singh[1, 2], Li You[2,3], Ankur Sikarwar[1, 2], Weixian Lei[4], Difei Gao[4],
Morgan B. Talbot[5, 6], Ying Sun[2], Mike Zheng Shou[4], Gabriel Kreiman[5], Mengmi Zhang[1, 2]
[1] Nanyang Technological University (NTU), Singapore [2] CFAR and I2R, Agency for Science, Technology and Research, Singapore,
[3] University of Wisconsin-Madison, USA, [4] Show Lab, National University of Singapore, Singapore,
[5] Boston Children's Hospital, Harvard Medical School, USA, [6] Harvard-MIT Health Sciences and Technology, MIT,
Address correspondence to mengmi@i2r.a-star.edu.sg

List of Supplementary Sections

List of Supplementary Figures

S1. Experiments with Human Subjects

S1.1. Psychophysics Experiments

We took three precautions to control data quality and ensure that subjects paid attention to the experiments.

1. Subjects had to click on randomly presented triangles during the training rounds, and their reaction times were recorded for attention checks.
2. Subjects had to recognize simple geometric shapes, such as 3D cubes, in randomly dispersed dummy trials during the testing rounds.
3. In each testing round trial, the "submit" button was disabled before the stimulus was shown for the full 200 millisecond presentation time to ensure that subjects were exposed to the stimulus.

For both MTurk and in-lab experiments, our results only incorporate data from subjects with 100% accuracy in recognition of geometric shapes.

S1.2. Mechanical Turk experiments

In our Amazon Mechanical Turk (MTurk) experiments, we collected responses from "master workers" with at least 1,000 approved human intelligence tasks (HITs) and a 95% approval rate. We collected responses from 242 subjects in total. After filtering subjects for data quality (**Sec S1.1**), we retained 169 subjects with 2-4 subjects for each tested curriculum.

In **Fig S1A**, we show the distribution of reaction times from the attention checks for all MTurk subjects. We show the accuracy histogram of subjects on attention check trials in **Fig S1B**.

In **Fig S2A** and **Fig S2B**, we show screenshots of the MTurk interface during the training and testing rounds of our experiment respectively. The exact same procedures and computer interfaces were used in the in-lab experiments.

We show the average accuracy of MTurk subjects over all tasks and an α vs. β (**Sec 3.3**) distribution for the Novel Object Dataset (NOD) in **Fig S4A** alongside results for in-lab subjects and Vanilla, EWC, and LwF continual learning (CL) algorithms. We also show the $\mathcal{F}$-scores of the top-5 vs. worst-5 performing curricula in **Fig S7A** alongside the best and worst curricula for in-lab subjects and the same CL algorithms. Overall, we observe a large effect of curriculum on learning performance in MTurk subjects. As shown in **Fig S7**, between the top-5 and worst-5 curricula for the MTurk experiments, $\mathcal{F}$ ranges from 1.82 ± 0.12 to 0.60 ± 0.09. This difference in $\mathcal{F}$ here is significant.

S1.3. In-lab Experiments

We augmented our study with in-lab experiments alongside the MTurk experiments to provide an additional layer of quality control. The exact same computer interfaces and experimental procedures were used for MTurk and in-lab experiments (**Fig S1**).

It was infeasible to access a pool of subjects large enough to exhaustively test all possible curricula in-lab. The in-lab experiments were conducted only on 6 curricula, 3 of which were among the top-5 curricula as determined in the MTurk experiments, and the other 3 of which were among the worst-5 curricula from the MTurk experiments. As shown in **Fig S6** (see legend for naming conventions), these 6 curricula are: ('fb3', 'fc1', 'fa1', 'fb1', 'fa2'), ('fb1', 'fc1', 'fb3', 'fa2', 'fa1'), ('fa1', 'fb3', 'fb1', 'fc1', 'fa2'), ('fa1', 'fa2', 'fc1', 'fb1', 'fb3'), ('fa1', 'fb3', 'fc1', 'fa2', 'fb1'), and ('fb1', 'fb3', 'fc1', 'fa2', 'fa1'). We recruited 60 subjects for in-lab experiments (10 for each curriculum), all of whom met the data quality criteria outlined in **Sec S1.1**.

We evaluated the $\mathcal{F}$ score (**Sec 3.3**) for each of the 6 in-lab curricula and compared $\mathcal{F}$ scores between in-lab and MTurk cohorts. As shown in **Fig S7**, between the top-3 and worst-3 curricula for the in-lab experiments, $\mathcal{F}$ ranges from 1.65 ± 0.19 to 1.30 ± 0.03. This difference in $\mathcal{F}$ aligns with our observations from the MTurk results, though unlike in the MTurk results the difference here is not statistically significant. Additionally, as can be seen from the curriculum visualizations in **Fig S6**, the best curricula from the MTurk experiments are not identical to the best curricula from the in-lab experiments. However, 2 out of the 3 top curricula from the in-lab experiments were among the top-5 curricula for MTurk subjects, and the second-worst curriculum for the in-lab subjects was also the second-worst for MTurk subjects.

S2. Additional Information on Datasets for Paradigm-II

We conducted our experiments using three datasets: MNIST [3], FashionMNIST [5], and CIFAR10 [2]. Each dataset consists of 10 object classes. If classes are learned one at a time, each curriculum is a permutation of 10 classes, resulting in more than $3e^6$ (10!) possible curricula per dataset. Running all possible curricula is not practical due to computational resource constraints. To mitigate this issue, we introduce two paradigms. In paradigm-I, we chose a subset of 5 classes for

each dataset (this paradigm produced the results described in the main paper, see **Sec** 3.1). In paradigm-II, we chose 5 tasks with 2 fixed classes each. In both paradigms, the order of the exemplars within each task is fixed and only the task sequence is permuted, resulting in a total of $5! = 120$ curricula. The pair-wise groupings of the 10-classes from each dataset for paradigm-II was as follows:
MNIST: (‘0,’ ‘1’), (‘2,’ ‘3’), (‘4,’ ‘5’), (‘6,’ ‘7’), (‘8,’ ‘9’).
FashionMNIST: (‘shirt,’ ‘sneaker’), (‘top,’ ‘trouser’), (‘bag,’ ‘boot’), (‘coat,’ ‘sandal’), (‘pullover,’ ‘dress’).
CIFAR10: (‘airplane,’ ‘automobile’), (‘frog,’ ‘horse’), (‘deer,’ ‘dog’), (‘ship,’ ‘truck’), (‘bird,’ ‘cat’)

S3. Analysis Across Experimental Settings

We explored whether empirical performance discrepancies among curricula were consistent across experimental settings, specifically the number of epochs, parameter initialization procedures, and learning rates.

For each experimental setting, we report the mean difference in curriculum discrepancy $\mathcal{H}$ (**Sec** 3.3) among all pairs of CL algorithms $\mathcal{A}$s (between-algorithm) and between $\mathcal{A}$s and the random curriculum designer (algorithm-random) on FashionMNIST (**Fig** S5). We vary only one experimental setting in each controlled experiment.

First, we varied the number of training epochs over 1, 10, and 20 per incremental step for all $\mathcal{A}$s. Curriculum discrepancy $\mathcal{H}$ was 0.16 lower on average in between-algorithms than in algorithm-random over all three CL algorithms (**Fig** S5**A**). This suggests that the relative efficacy of different curricula is similar regardless of whether algorithms train for one or multiple epochs.

Next, we vary the learning rates of all CL algorithms over $0.5e^{-3}$, $1e^{-3}$, and $2e^{-3}$. We observe $\mathcal{H}$ values that are lower by 0.02 on average in between-algorithms comparisons than in algorithm-random comparisons (**Fig** S5**B**). However, at the highest learning rate of $2e^{-3}$, the difference is much smaller than at lower learning rates. This suggests the hypothesis that, at high learning rates, curriculum effects may be either less impactful or less consistent in terms of which curricula are optimal.

Lastly, we tried several different network parameter initialization procedures: Gaussian, Uniform, and Xavier [1]. We observed an average decrease of 0.03 in the curriculum discrepancy from algorithm-random to between-algorithms (**Fig** S5**C**). However, this decrease is much smaller for Xavier initialization than for the other two initialization procedures, suggesting that the extent to which optimal curricula agree across CL algorithms is dependent on the choice of parameter initialization procedure in at least some cases.

S4. Curriculum Affects Learning Performance Across Algorithms, Datasets, and Paradigms

We analyze curriculum effects for three continual learning algorithms (**Sec** 3.2) on three image datasets in both paradigm-I and paradigm-II (**Sec** 3.1). For each analysis, we provide α versus β plots (**Fig** **S14**-**S21**), and the $\mathcal{F}$ distribution for the top-10 and bottom-10 curricula (**Fig** S23, S24). Overall, the results suggest that curriculum significantly impacts performance in online class-incremental CL. Across all 18 scenarios (3 CL algorithms × 3 datasets × 2 paradigms), we observe statistically significant differences in performance between the 10 best and 10 worst curricula.

S5. Learning Effectiveness $\mathcal{F}$ as a Function of Time

We present the task-wise $\mathcal{F}$ score (**Sec** 3.3) of the Vanilla CL algorithm, a “random” model, and an “overfitting” model (**Fig** S29) across three datasets for paradigm-I (**Sec** 3.3). In each task, the random model makes a random guess of the class label out of all the learned classes. The theoretical over-fitting model has perfect accuracy on the current task but has 100% catastrophic forgetting and 0% accuracy on previous tasks. We observe that the variance of $\mathcal{F}$ increases with increasing task number, implying a stronger curriculum effect with longer task sequences. We also observe that, even for the Vanilla algorithm, an effective curriculum leads to higher $\mathcal{F}$ than the overfitting and random models. Note that the overfitting model completely forgets task 1 when learning task 2; thus, $\mathcal{F}_{T=2} = 2/(1-0+1/0.5) = 0.67$ which is less than chance prediction. In case of chance, since each class would be assigned equal probability, we would have $\mathcal{F}_{T=2} = 2/(1-0.5+1/0.5) = 0.8$.

S6. Alternative Curriculum Ranking Agreement Metric: Spearman’s Rank Correlation Coefficient

As referenced in **Sec** 3.3, we also calculate Spearman’s rank correlation coefficients for curriculum ranking agreements on MNIST in paradigm-I (**Sec** 3.1), showing that it leads to the same conclusions as those reached using $\mathcal{H}$. We calculated Spearman’s correlation coefficients of 0.26, 0.08, and 0.0002 for between-algorithms, algorithm-CD, and algorithm-random comparisons for MNIST in paradigm-I (averaging among pairs of CL algorithms As). These findings are consistent with those in **Sec 5.4** based on $\mathcal{H}$: CL algorithms agree to a significant extent on empirical rankings of curricula, and our CD predicts these empirical rankings better than a random CD.

S7. Our CD Predicts Optimal Curricula in Paradigm-II Based on Recall@K Measurements

Following the same figure interpretation as for paradigm-I in **Fig 4**, we report Recall@K results for paradigm-II in **Fig S27**. We found that our CD predicted optimal curricula more accurately than the random model on average across all three datasets, particularly at larger values of k. Moreover, we see no clear evidence that the performance of our CD is dependent on the difficulty of the classification tasks to be learned, since it performs well across three datasets with varying complexity.

S8. Analysis of Curriculum Discrepancy in Paradigm-II

Fig S22 illustrates the discrepancy $\mathcal{H}$ between curriculum rankings determined empirically by CL algorithms, heuristically by our curriculum designer (CD), and randomly by the random curriculum designer on MNIST, FashionMNIST, and CIFAR10 (**Sec 3.1**) in paradigm-II (10 classes arranged in 5 binary tasks, **Sec 3.1**). A decrease in $\mathcal{H}$ indicates an increase in the agreement between curriculum rankings (**Sec 3.3**).

Like in Paradigm-I, we conclude that CL algorithms share a comparable set of top-ranked curricula across three datasets in Paradigm-II. We also assess curriculum agreement between our CD and CL algorithms. We observe an decrease of 0.01 in the discrepancy from algorithm-random to algorithm-CD in CIFAR10. However, our CD fails for MNIST and FashionMNIST, yielding higher curriculum discrepancy with empirically ranked curricula than a random CD, despite identifying optimal curricula better than a random CD according to Recall@K (**Sec S7, Fig S27**). This suggests that although our CD identified the highest-performing curricula relatively well for MNIST and FashionMNIST in this setting, it did not accurately predict the rankings of less effective curricula further down in the rankings. In any case, there is still a great deal of room for improvement in predicting optimal curricula across datasets, algorithms, and training regimens.

S9. CD Ablation Study in Paradigm-II

We report the effects of ablating several CD design decisions in Paradigm-I in **Fig 5**, and repeat them in **Fig S28A** for convenience. **Fig S28B** shows CD ablation results for paradigm-II. We follow the same figure conventions as **Fig 5**. Unlike in the results from paradigm-I (see **Sec 5.3**), we did not observe clear benefits of our specific CD design choices in paradigm-II (e.g., as indicated by zero recall at k=5).

S10. Curriculum Influences Performance of a Naive Replay Algorithm in Class-Incremental Online CL

To extend our study of class-incremental online CL with Vanilla, EWC, and LwF, we investigate the effects of curricula on a naive replay CL algorithm. This algorithm used a replay buffer size equivalent to 10% of the training set of each task (for example, if the training set comprised x images per task, the buffer size would $0.1x$) and adopted a random sampling strategy to select samples for the memory buffer. We did not experiment with the ordering of the replayed examples themselves.

As observed in **Fig S25**, for MNIST, the average $\mathcal{F}$ scores ($\pm$ standard deviation) were 1.55 ± 0.06 and 0.93 ± 0.04 for the top-10 worst-10 curricula respectively. For FashionMNIST the average $\mathcal{F}$ scores were 1.26 ± 0.07 and 0.79 ± 0.04, and for CIFAR10 they were 1.14 ± 0.04 and 0.63 ± 0.04. This suggests that curriculum plays a crucial role in the performance of replay-based continual learning algorithms. Across all three datasets, we observed that the top curricula outperform the worst curricula significantly.

We also assessed the curriculum discrepancy $\mathcal{H}$ (**Fig S25**) between pairs of curriculum rankings determined by CL algorithms (Vanilla, EWC, LwF and naive-replay; accounting for all pairs of CL algorithms) including the naive-replay CL algorithm, and a random curriculum designer. We observe a statistically significant decrease in $\mathcal{H}$ from 0.60 ± 0.001 to 0.40 ± 0.07 in $\mathcal{H}$ from between-algorithms to algorithm-random for MNIST. For FashionMNIST, we observe a statistically significant $\mathcal{H}$ decrease from 0.60 ± 0.001 to 0.40 ± 0.06, and for CIFAR10 we observe a statistically significant $\mathcal{H}$ decrease from 0.62 ± 0.002 to 0.38 ± 0.09. This implies that, compared to the agreement between the curricula ranked randomly and curricula ranked empirically by CL algorithms, the CL algorithms including naive-replay share comparable curriculum rankings.

S11. Curriculum Influences Performance in Offline Class-Incremental Learning

We extend our investigation of curriculum effects in CL to offline class-incremental CL, where multiple passes over the data within each task are allowed. **Fig** S23 highlights the effect of curricula on the Vanilla, EWC and LwF algorithms (**Sec** 3.2) over three datasets (**Sec** 3.1) in this offline CL setting. Despite multi-epoch training on each task, the results are consistent with our findings as highlighted in **Sec** 5.1 and **Sec** S4.

S12. Curriculum Strongly Affects Performance in Continual Visual Question Answering

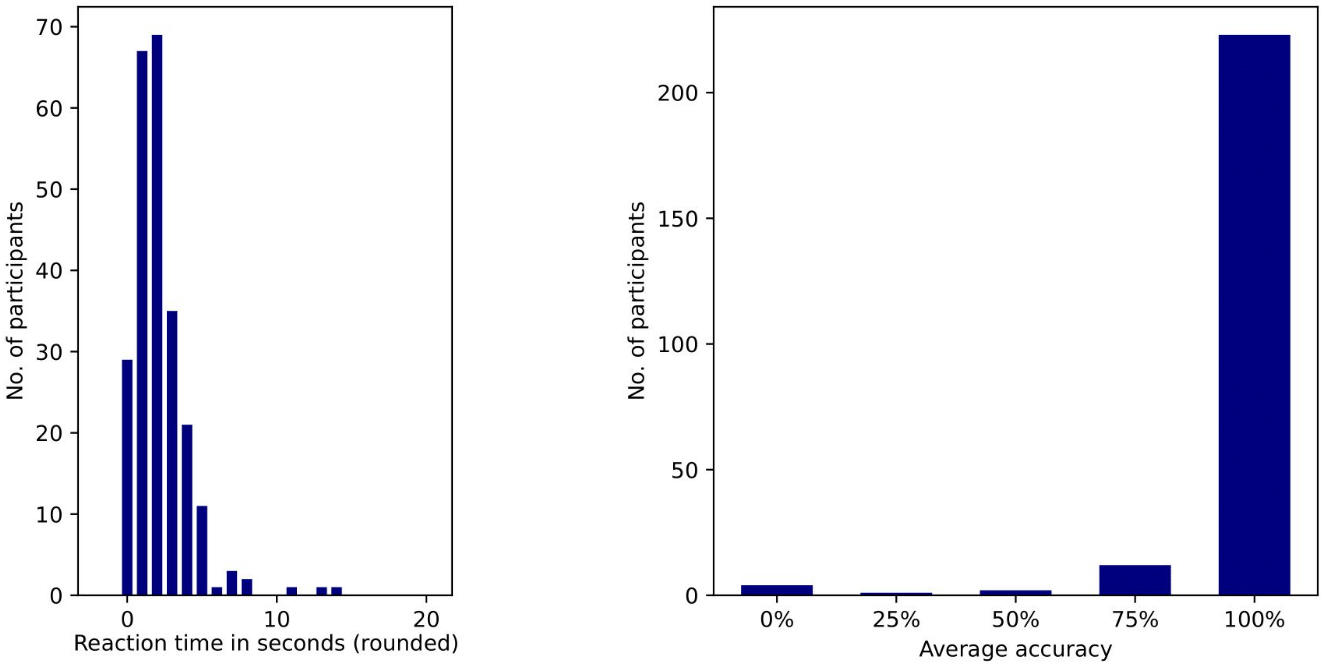
To study the impact of curriculum in a multi-modal setting, we conducted additional experiments using Vanilla and EWC CL algorithms on the CLOVE VQA dataset ([4]). CLOVE is a benchmark dataset for CL in a VQA setting, and comprises question-answer (QA) pairs in five groups for function-incremental settings. The QA pairs are categorized based on the five functions of knowledge reasoning, object recognition, attribute recognition, relation reasoning, and logic reasoning. Since computing the results across all possible curricula ($5! = 120$) was infeasible due to limited computational resources, we sampled 16 curricula at random. Despite only sampling a small subset of possible curricula, in **Fig** S26 we observe strong curriculum effects in the function incremental setting. $\mathcal{F}$ between the top-5 sampled curricula and worst-5 sampled curricula ranges from 0.64 ± 0.05 to 0.36 ± 0.02 using the Vanilla algorithm and ranges from 0.64 ± 0.04 to 0.35 ± 0.02 using the EWC algorithm

We assessed the curriculum discrepancy $\mathcal{H}$ (**Fig** S26) between pairs of curriculum rankings determined empirically by the Vanilla and EWC algorithms, and by a random curriculum designer. We observe a significant decrease in $\mathcal{H}$ from 0.76 ± 0.0004 to 0.24 ± 0.0001 from algorithm-random to between-algorithms, indicating that the CL algorithms share a comparable set of top-ranked curricula when compared to the agreement between randomly and empirically ranked curricula.

It is intriguing that, in general, the curriculum effects we observe in VQA are dramatically larger than those we observe in image classification. This experiment further supports the conclusion that curriculum plays an important role in continual learning, perhaps especially in complex continual learning settings such as continual VQA.

S13. Statistical Analysis

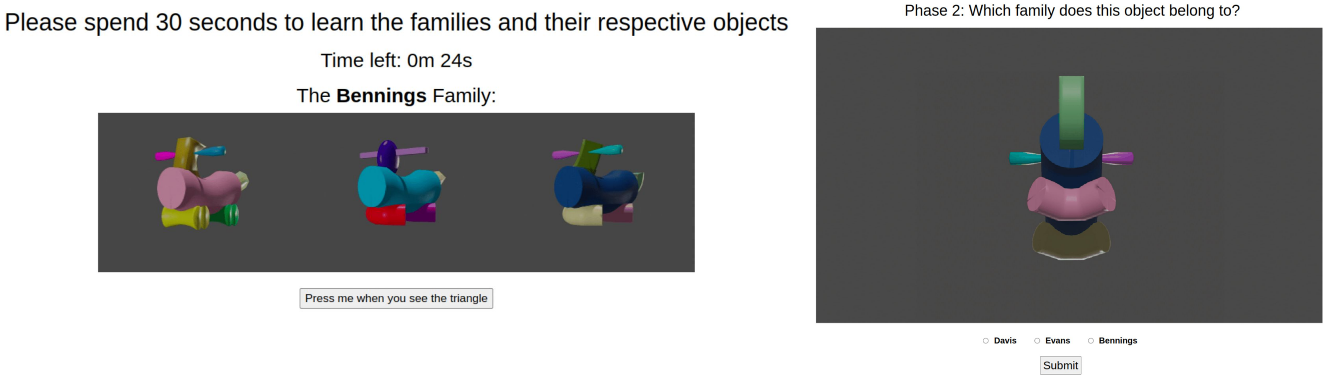
We employed two-sample t-tests to compute statistical significance in the following cases: (1) comparing the top-k $\mathcal{F}$ scores to the bottom-k $\mathcal{F}$ scores to establish the presence of curriculum effects (see **Sec** S1, S4, S12, S10 and **Fig** S7, S23, S24, S25, S26), and (2) comparing two sets of $\mathcal{H}$ (**Fig** S25, S26) to discern if the curriculum agreement between two distributions varies significantly or not. We use the asterisk symbol * in all relevant figures to denote significant p-values ($p < 0.05$) in 2-sample t-tests, and use "n.s." to denote higher non-significant p-values. Errorbars are also presented to indicate standard deviation across all test trials.



(A)

(B)

Figure S1: **Reaction time and attention check accuracy histograms for MTurk experiments.** (A) Reaction time distribution for all subjects in attention checks during training rounds. subjects were required to click on randomly presented triangles during the training rounds and their reaction time was recorded. (B) Average accuracy of all subjects on attention checks during testing rounds. We only used data from subjects who satisfied the criteria delineated in (**Sec S1**).



Please spend 30 seconds to learn the families and their respective objects

Time left: 0m 24s

The **Bennings** Family:

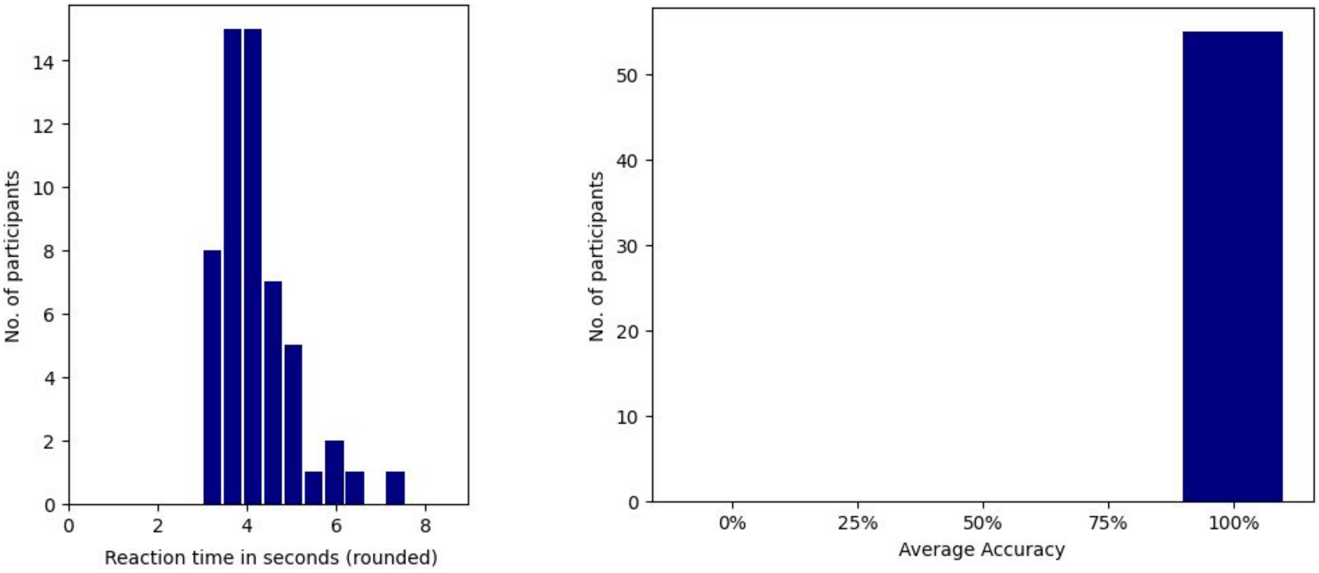
Phase 2: Which family does this object belong to?

○ Davis ○ Evans ○ Bennings

(A)

(B)

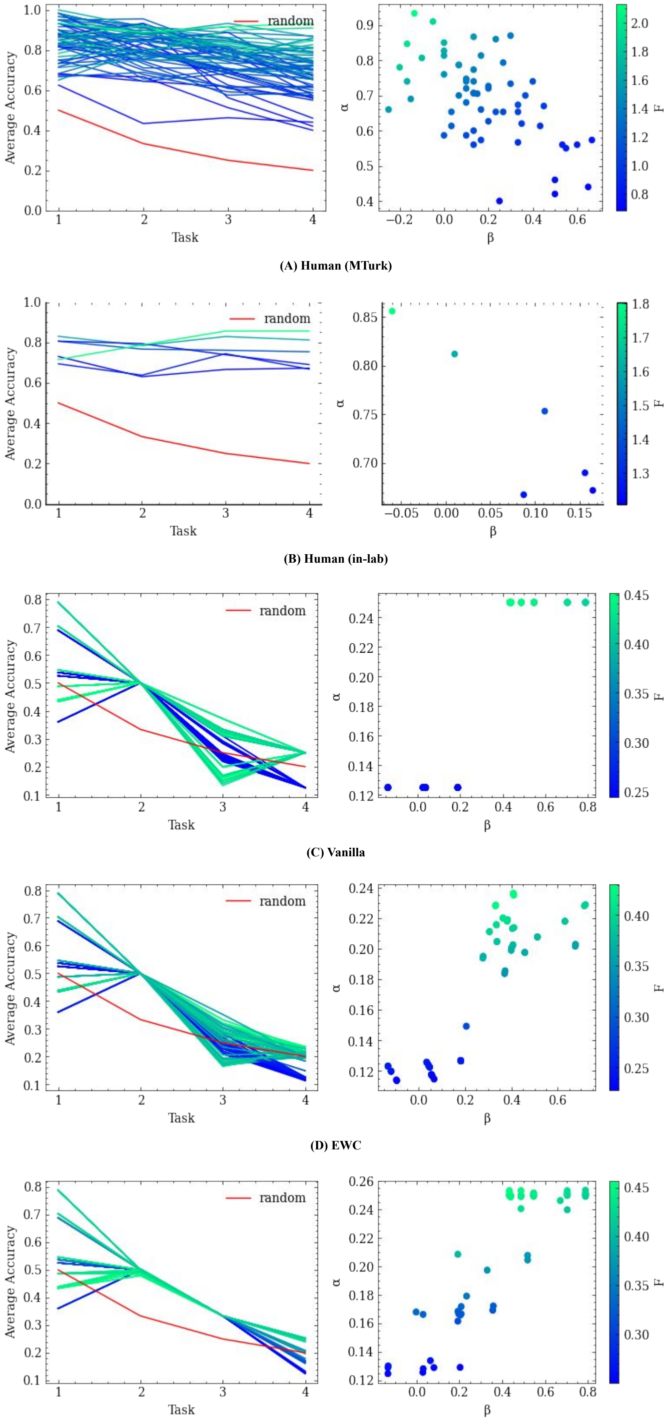
Figure S2: **MTurk interface schematics.** Screenshots of the MTurk interface during the training rounds (A) and testing rounds (B).



(A)

(B)

Figure S3: **Reaction time and attention check accuracy for in-lab experiments.** (A) Reaction time distribution for all subjects in attention checks during training rounds. Subjects were required to click on randomly presented triangles during the training rounds and their reaction time was recorded. On the x-axis, we show the reaction time in seconds (rounded). (B) Average accuracy of all subjects in attention checks during testing rounds. All in-lab subjects were included in our analysis, since all subjects' data satisfied the inclusion criteria in **Sec S1**.



(A) Human (MTurk)

(B) Human (in-lab)

(C) Vanilla

(D) EWC

(E) LwF

Figure S4: **Curriculum effects on performance on NOD for humans and for Vanilla, EWC and LwF CL algorithms (Sec 3.2, Sec 3.4).** We report the average accuracy across all tasks in the left-hand panel for each condition. We also plot α vs β (**Sec 3.3**) in the right-hand panel for each condition. The effectiveness measure $\mathcal{F}$ (**Sec 3.3**) incorporates both α and β.

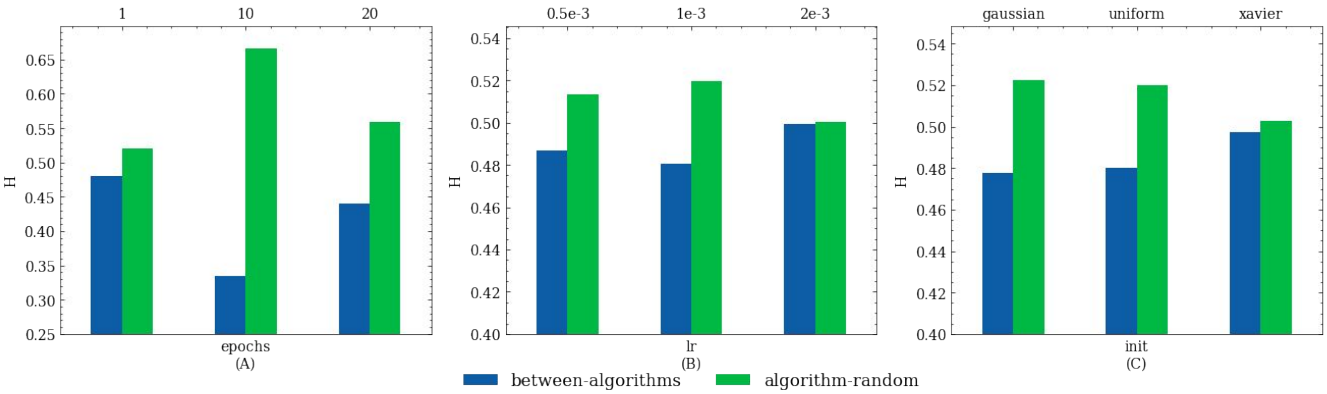
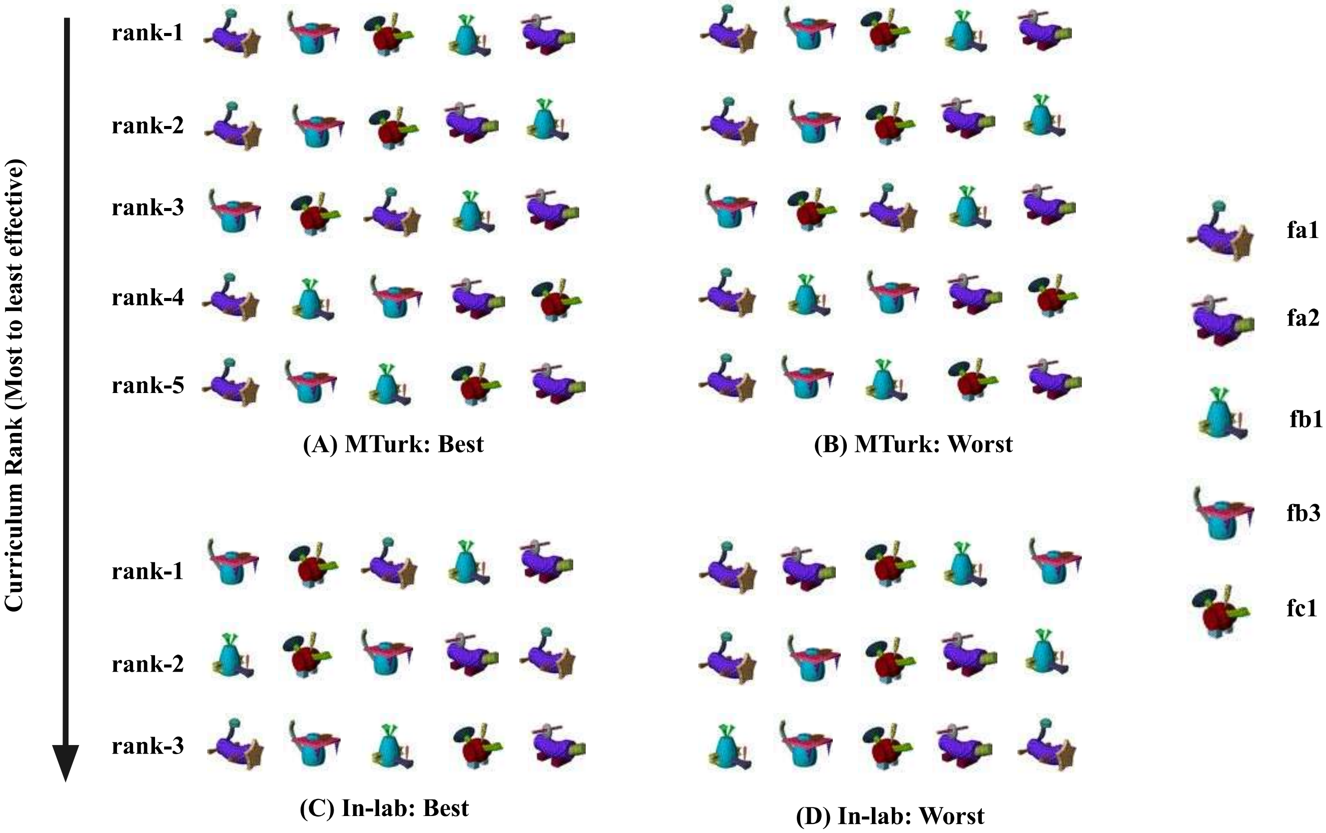


Figure S5: **Curriculum agreement (low curriculum discrepancy $\mathcal{H}$) among CL algorithms persists across different experimental settings on FashionMNIST.** The discrepancy between two sets of ranked curricula is measured as $\mathcal{H}$, with smaller values indicating lower discrepancy and higher curriculum agreement (**Sec 3.3**). Within each pair of bars, the discrepancy between pairs of ranked curriculum sets for CL algorithms $\mathcal{A}$s (between-algorithms) is presented on the left (blue), and that between $\mathcal{A}$ and the randomly ranked curricula (algorithm-random) is on the right (green). We vary the number of epochs (A), the learning rates (lr) (B), and the network parameter initialization procedure (C). For visualization purposes, within each pair of bars, we normalize the $\mathcal{H}$ value over between-algorithm and algorithm-random so that the sum of these two discrepancy values (green + blue) always equals 1. Normalization does not alter the main conclusion that curriculum discrepancy is always lower in the between-algorithms condition, meaning the same curricula work well (and the same curricula work poorly) across a range of experimental conditions.



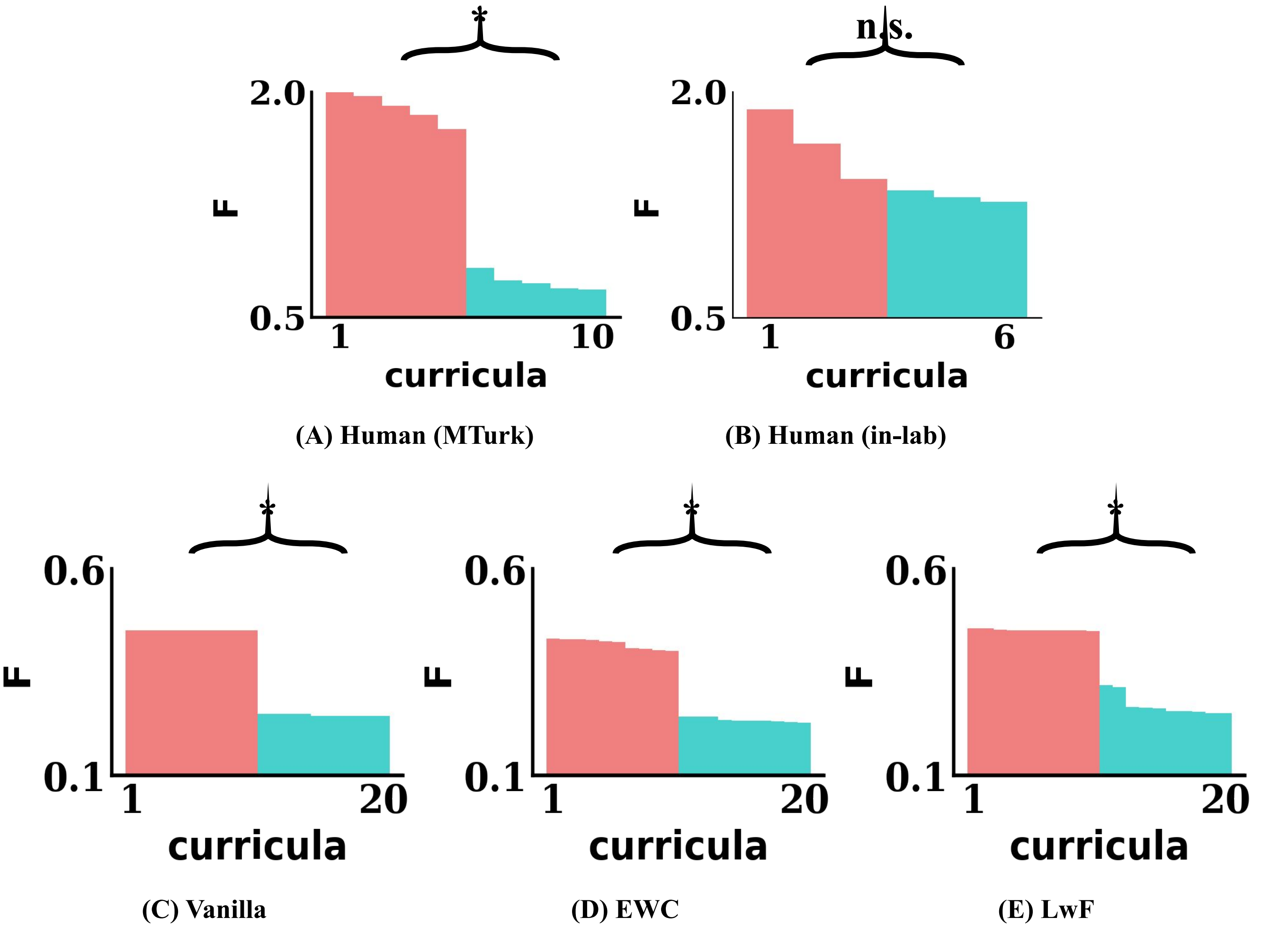
(A) MTurk: Best

(B) MTurk: Worst

(C) In-lab: Best

(D) In-lab: Worst

Figure S6: **Experimentally determined best and worst curricula on NOD (Sec 3.4) for MTurk (A-B, Sec 3.4) and in-lab (C-D, Sec S1) human subjects.** Each row in the figure is one curriculum. The curricula are arranged from best to worst with the best curricula at the top.



(A) Human (MTurk)

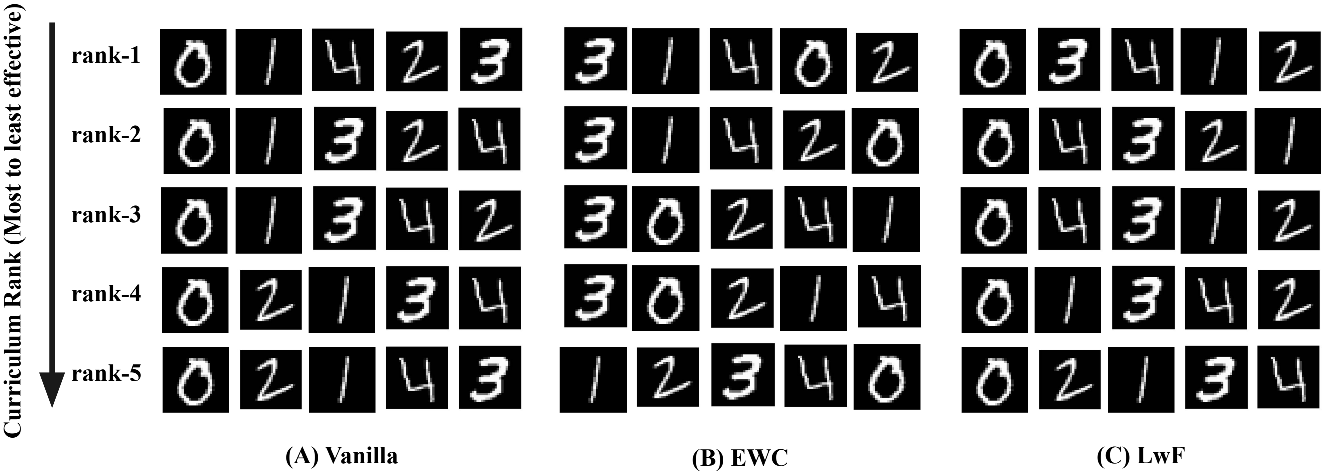
(B) Human (in-lab)

(C) Vanilla

(D) EWC

(E) LwF

Figure S7: **Best and worst k curricula on NOD (Sec 3.1) for (A) MTurk human subjects (top 5 vs bottom 5), (B) in-lab human subjects (top 3 vs bottom 3), (C) Vanilla (top 10 vs bottom 10), (D) EWC (top 10 vs bottom 10), and (E) LwF (top 10 vs bottom 10)**. The plot shows the $\mathcal{F}$-scores for the best curricula (red) and the worst curricula (blue) as well as the statistical significance (* = statistically significant) determined via two-sample t-tests on the $\mathcal{F}$-scores of the best and worst curricula.

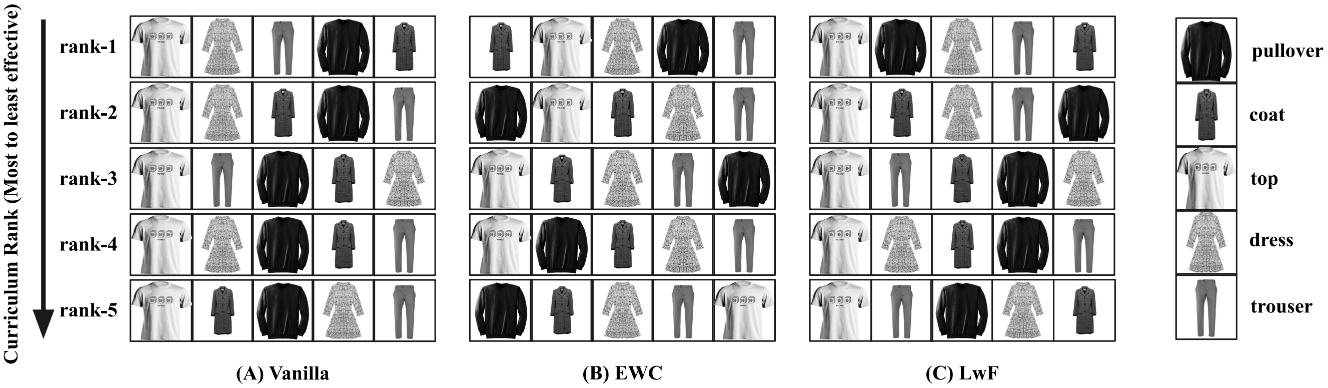


(A) Vanilla

(B) EWC

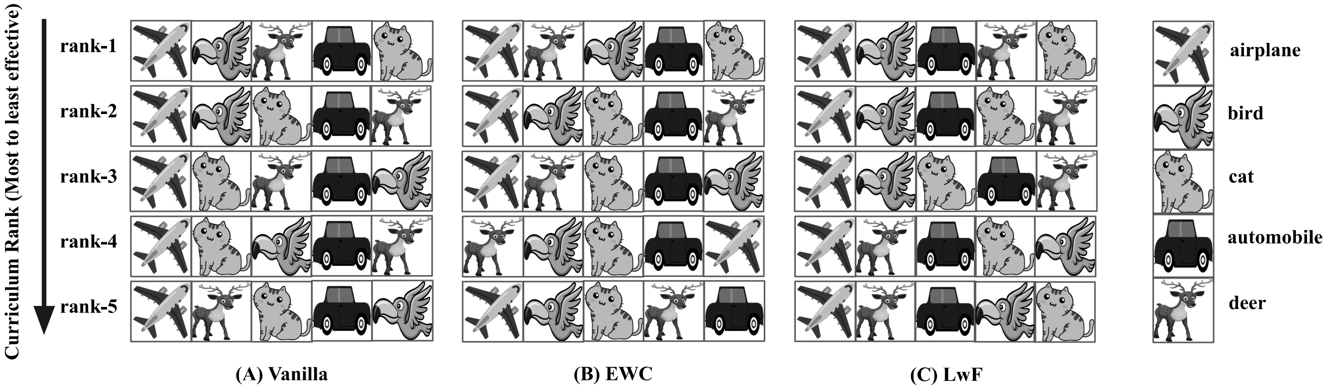
(C) LwF

Figure S8: **Empirically determined top-5 curricula on MNIST for Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top.



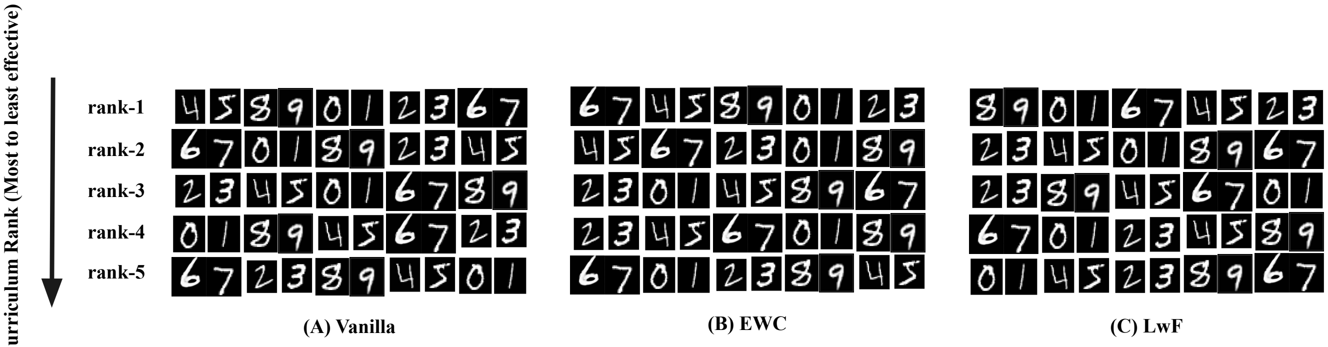
(A) Vanilla **(B) EWC** **(C) LwF**

Figure S9: **Empirically determined top-5 curricula on FashionMNIST for Vanilla, EWC, and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top.



(A) Vanilla **(B) EWC** **(C) LwF**

Figure S10: **Empirically determined top-5 curricula on CIFAR10 for Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top. For ease of interpretation, cartoon images are used to represent each class instead of actual CIFAR10 images.



(A) Vanilla **(B) EWC** **(C) LwF**

Figure S11: **Empirically determined top-5 curricula on MNIST for Vanilla, EWC, and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top.

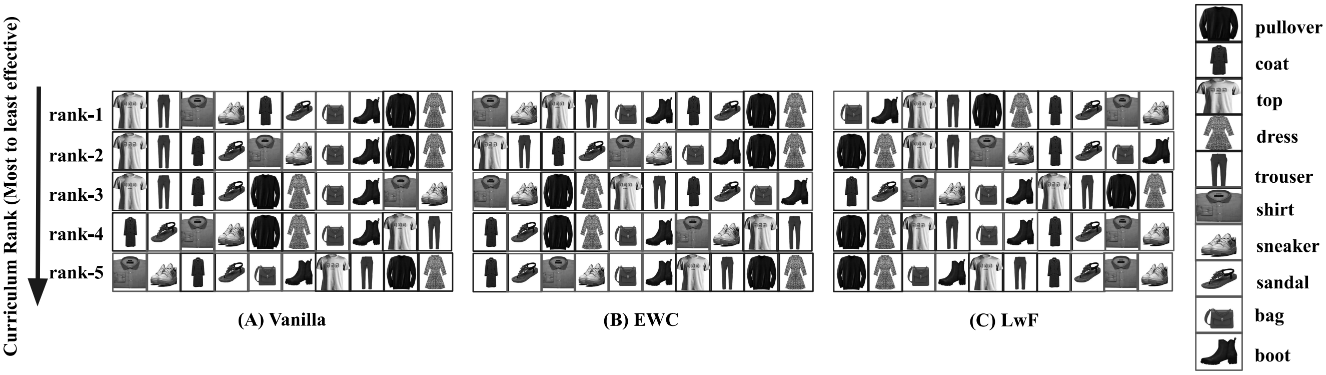


Figure S12: **Empirically determined top-5 curricula on FashionMNIST for Vanilla, EWC, and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top.

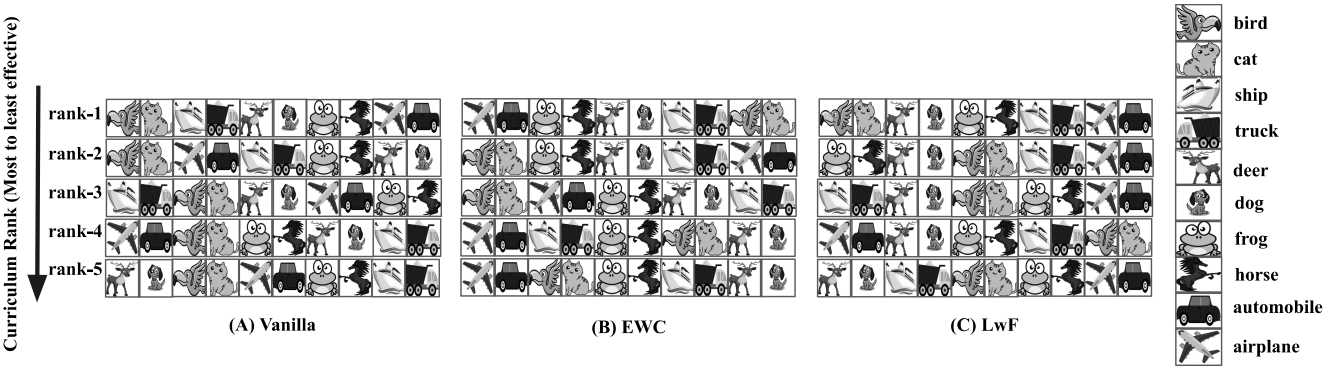
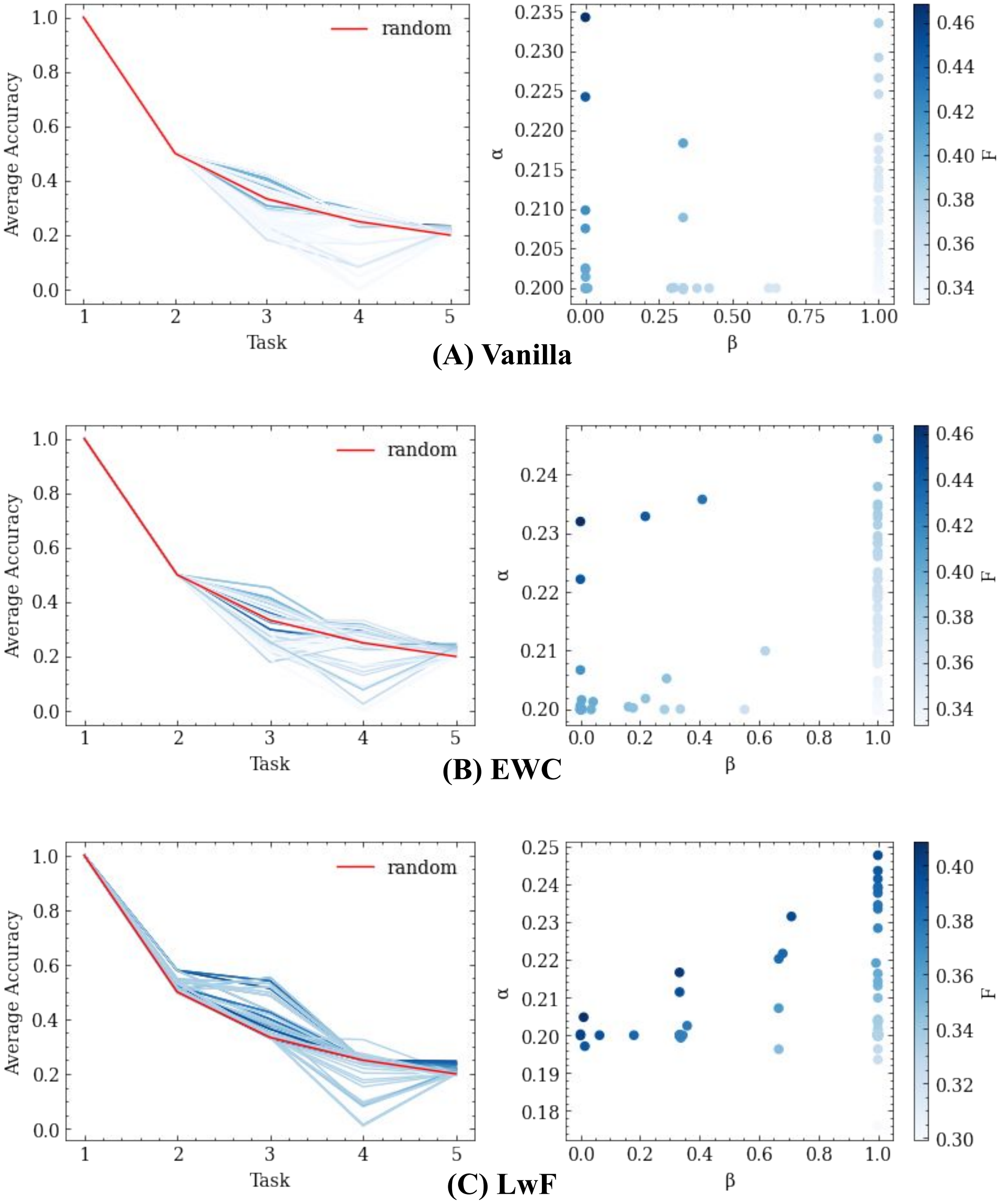


Figure S13: **Empirically determined top-5 curricula on CIFAR10 for Vanilla, EWC, and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** Each row in the figure is one curriculum. Curricula are in descending order of effectiveness, with the best curriculum at the top. For ease of interpretation, cartoon images are used to represent each class instead of actual CIFAR10 images.

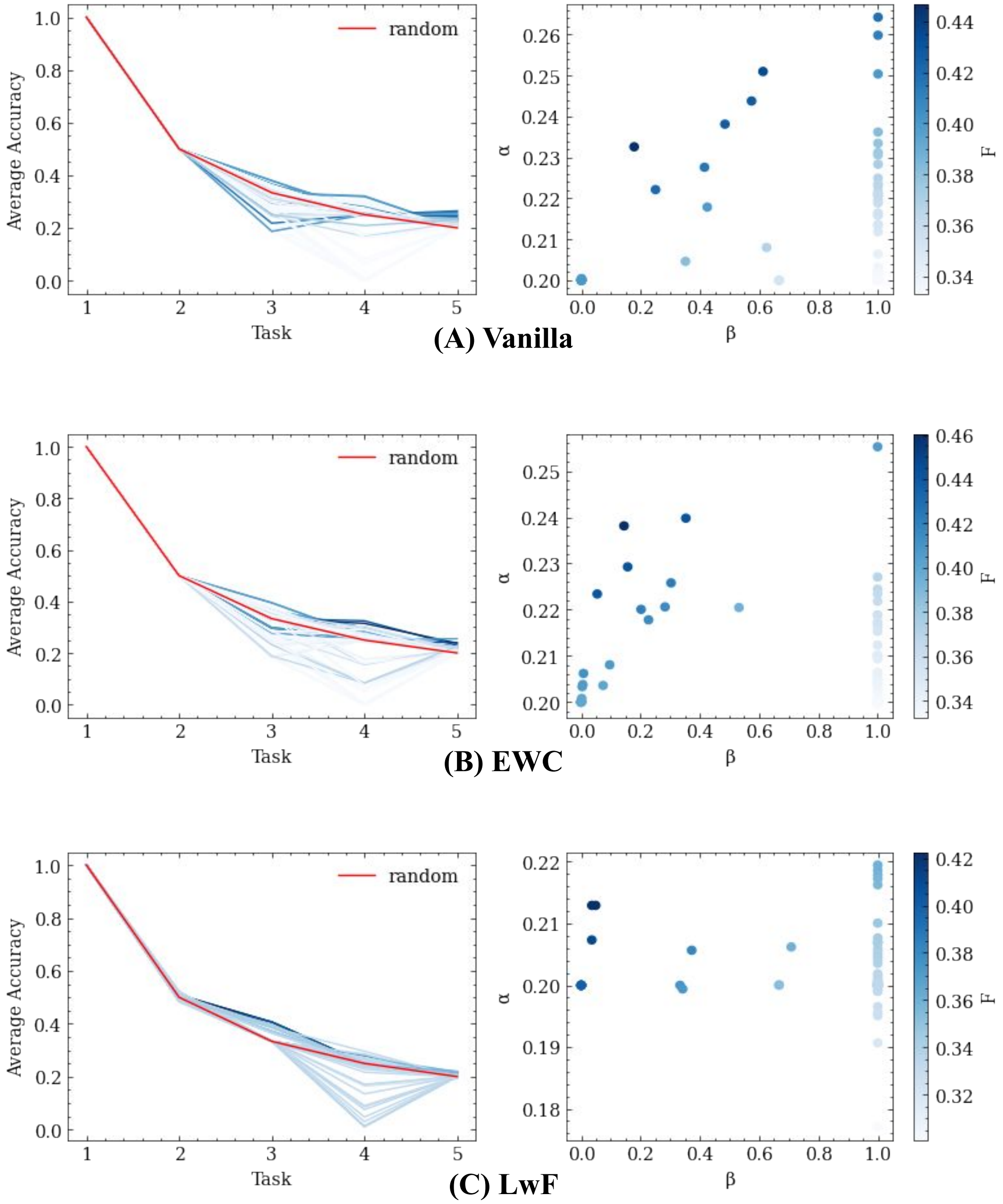


(A) Vanilla

(B) EWC

(C) LwF

Figure S14: **Curriculum affects performance on MNIST for the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.

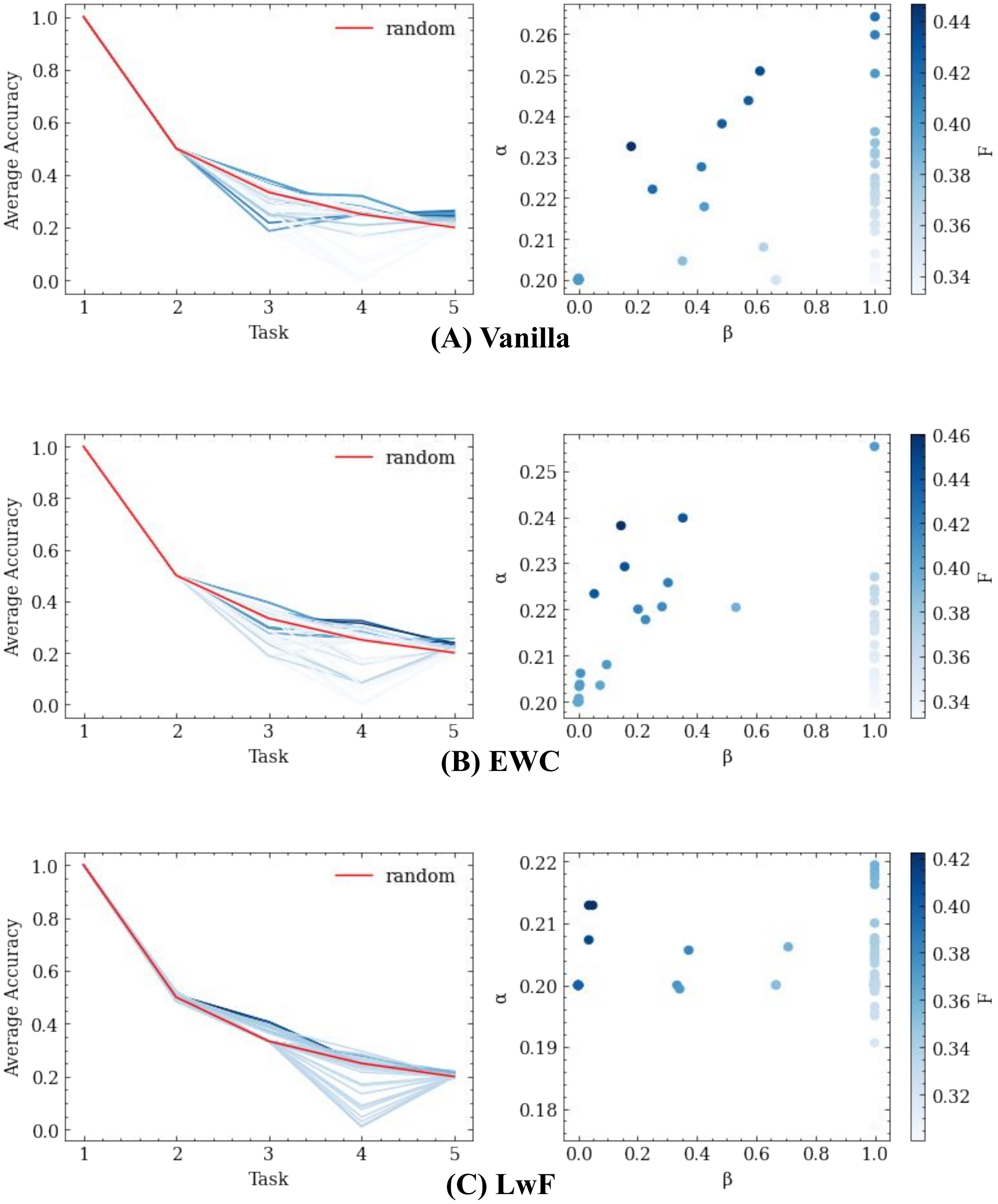


(A) Vanilla

(B) EWC

(C) LwF

Figure S15: **Curriculum affects performance on FashionMNIST of the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.



(A) Vanilla

(B) EWC

(C) LwF

Figure S16: **Curriculum affects performance on CIFAR10 of the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.

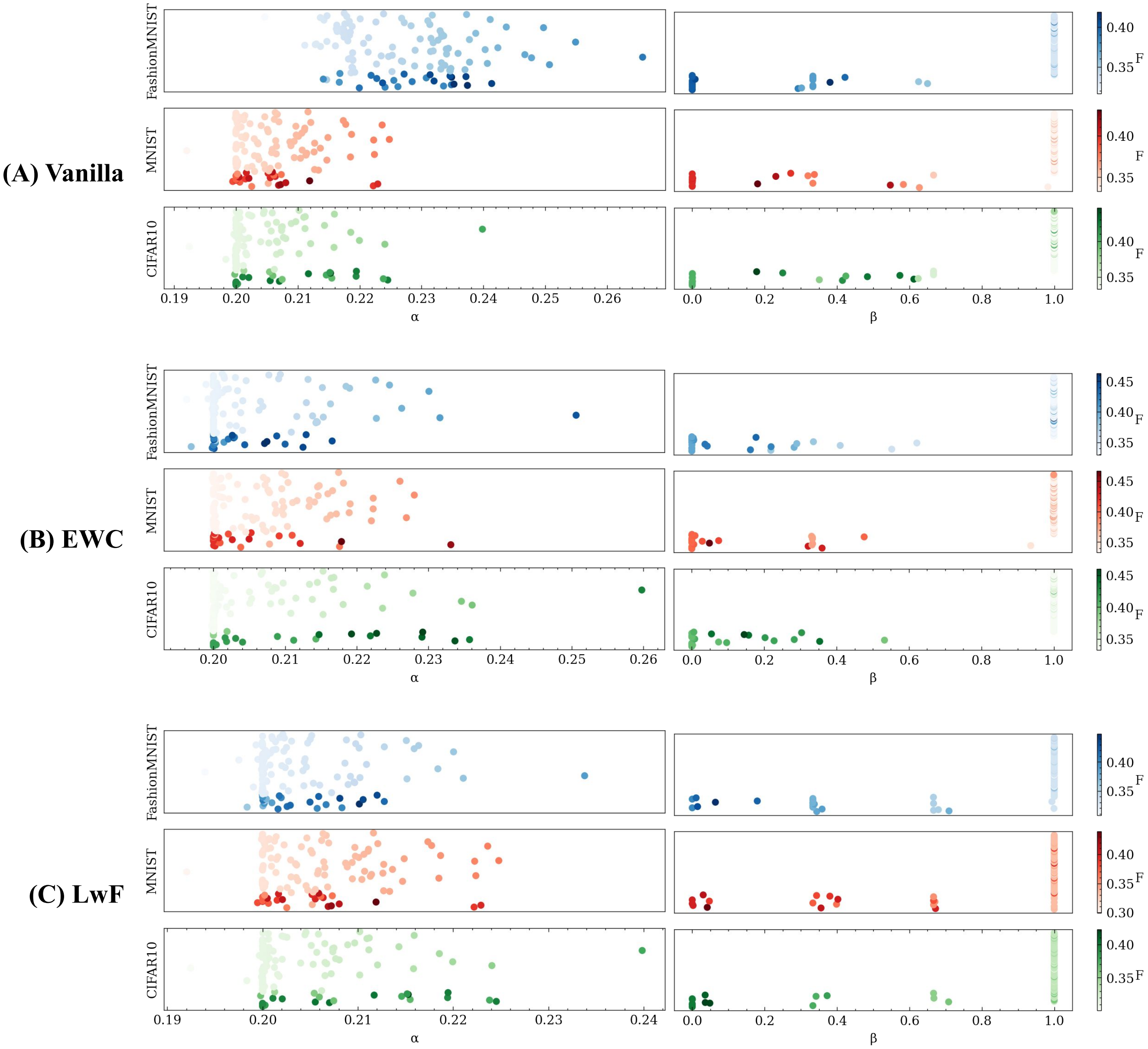
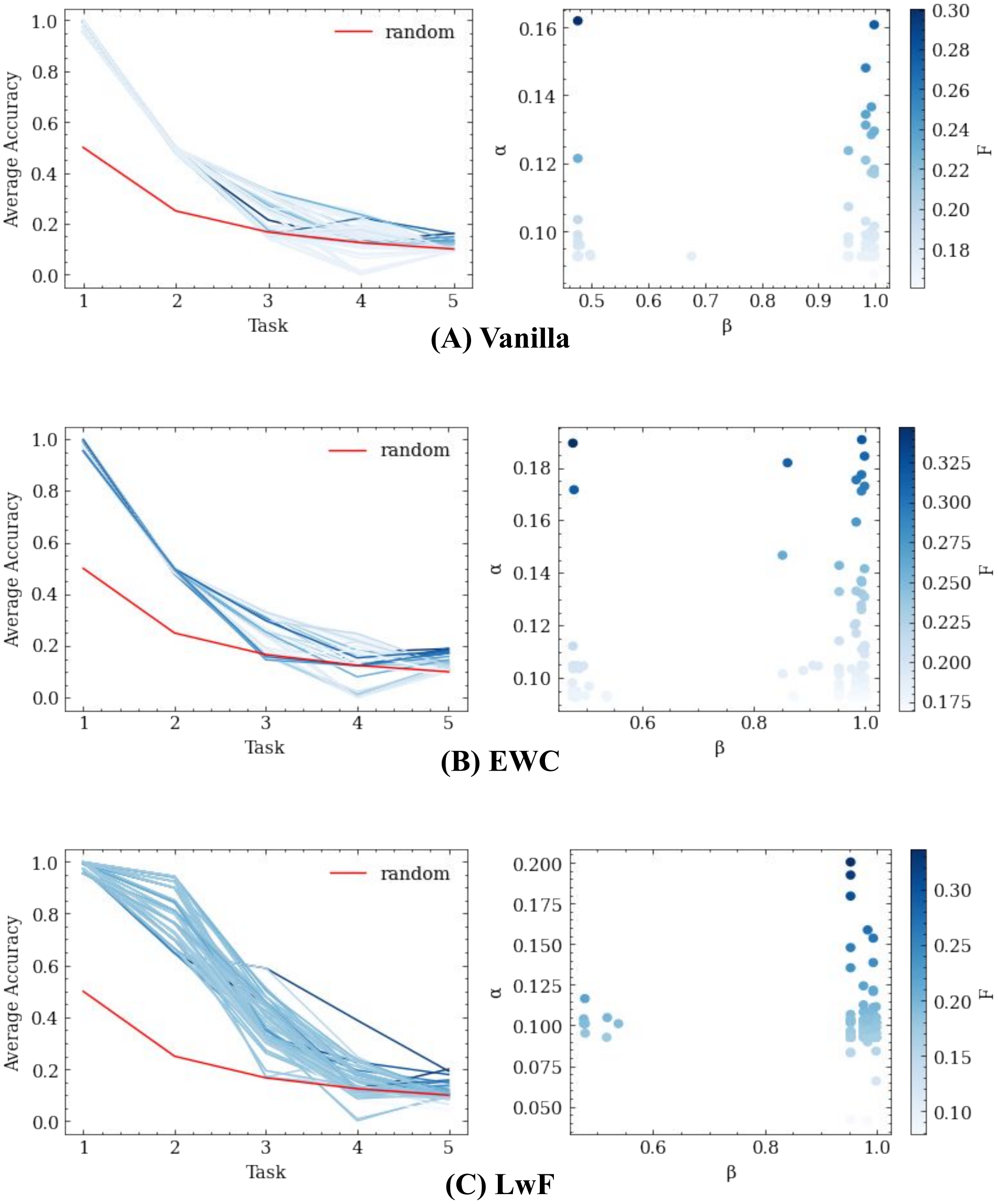


Figure S17: **Curriculum affects learning performance of the (A) Vanilla, (B) EWC, and (C) LwF CL algorithms (Sec 3.2) across three datasets: MNIST, FashionMNIST, and CIFAR10 (Sec 3.1) in paradigm-I (5 classes, Sec 3.1).** Note that the y-axis does not carry any meaning. All the dots are randomly spread along the y-axis for easy visualization of the α and β distributions. This figure uses the same design conventions as **Fig 2**.

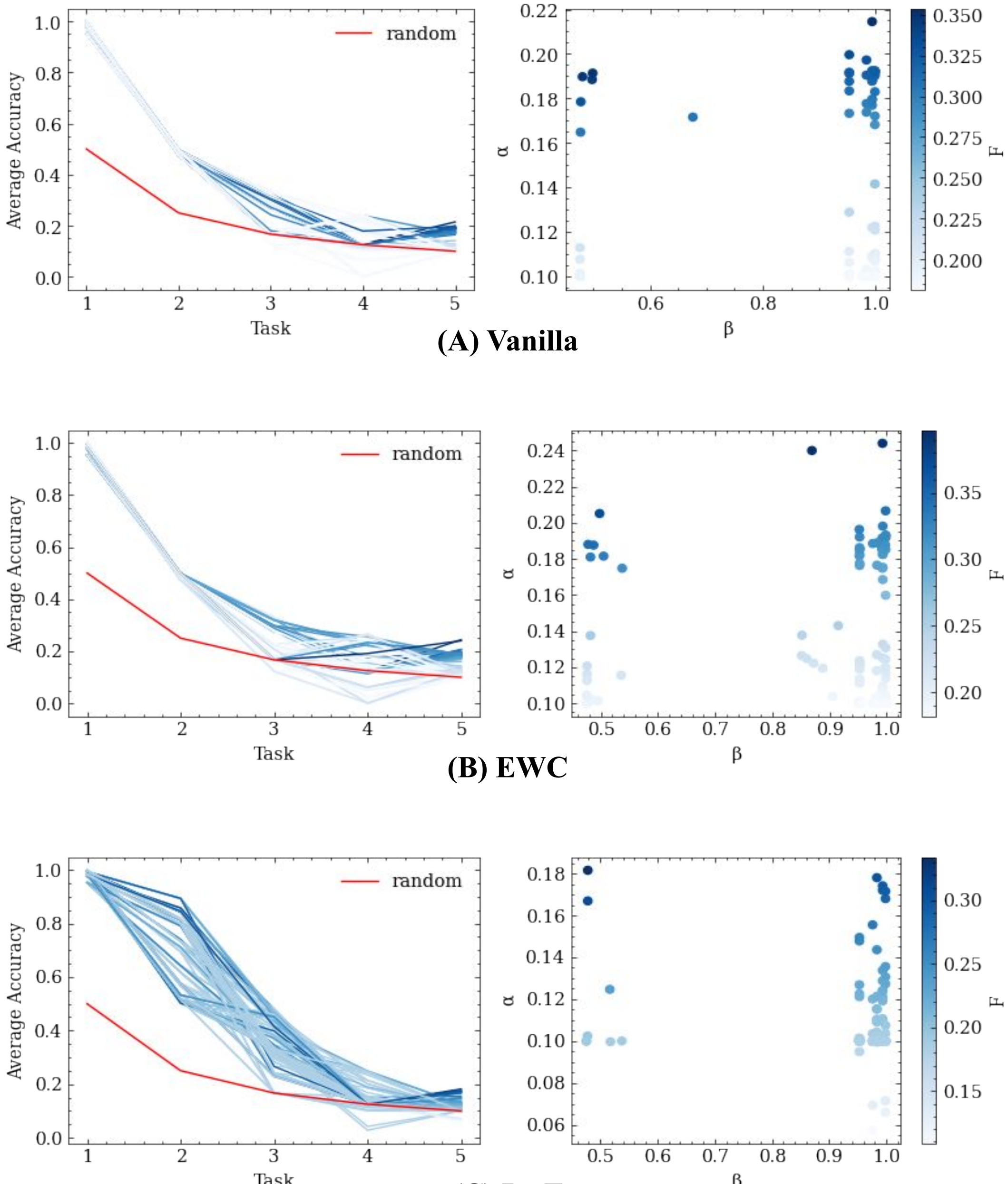


(A) Vanilla

(B) EWC

(C) LwF

Figure S18: **Curriculum affects performance on MNIST of the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.

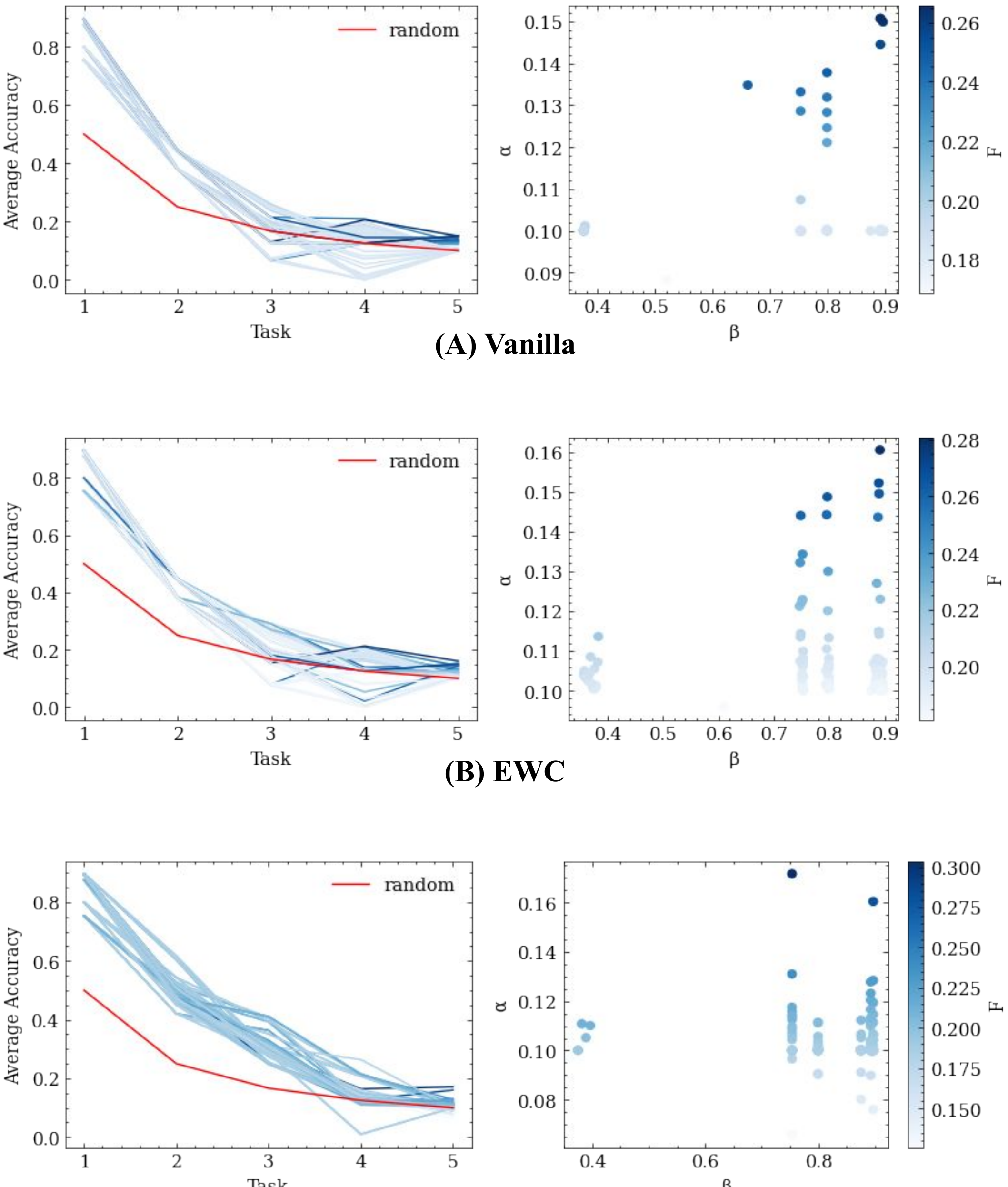


(A) Vanilla

(B) EWC

(C) LwF

Figure S19: **Curriculum affects performance on FashionMNIST of the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.



(A) Vanilla

(B) EWC

(C) LwF

Figure S20: **Curriculum affects performance on CIFAR10 of the Vanilla, EWC and LwF CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** This figure follows the same design conventions as **Fig S4**.

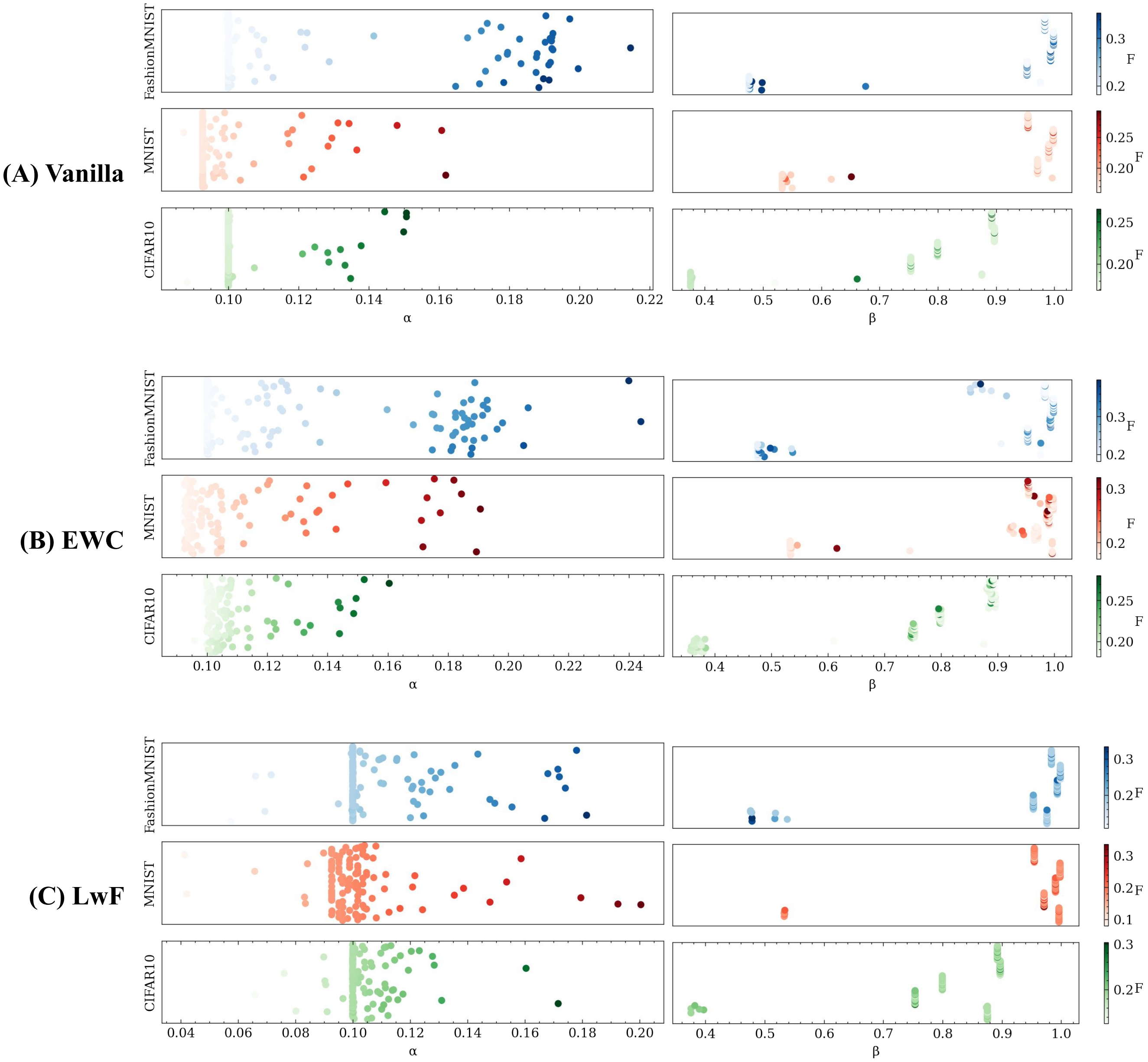


Figure S21: **Curriculum affects learning performance of the (A) Vanilla, (B) EWC, and (C) LwF CL algorithms (Sec 3.2) across three datasets: MNIST, FashionMNIST, and CIFAR10 (Sec 3.1) in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** Note that the y-axis does not carry any meaning. All the dots are randomly spread along the y-axis for easy visualization of the α and β distributions. This figure uses the same design conventions as **Fig 2**.

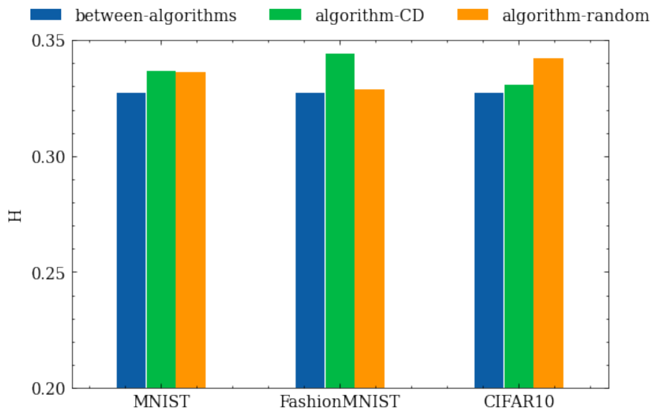


Figure S22: **Like in paradigm-I, in paradigm-II there is agreement among methods in ranking curricula by effectiveness.** Different CL algorithms agree with each other on which curricula are more effective than others, and also with our CD's heuristic estimates of relative curriculum optimality. Curriculum discrepancy $\mathcal{H}$ (**Sec 3.3**) is reported between pairs of CL algorithms (between-algorithms, blue, averaging across all pairs), between CL algorithms and our CD (algorithm-CD, green, averaging across CL algorithms), and between CL algorithms and the random CD (algorithm-random, orange, averaging across CL algorithms) across MNIST, FashionMNIST, and CIFAR10 (**Sec S8**). See **Sec S8** for an analysis of these results.

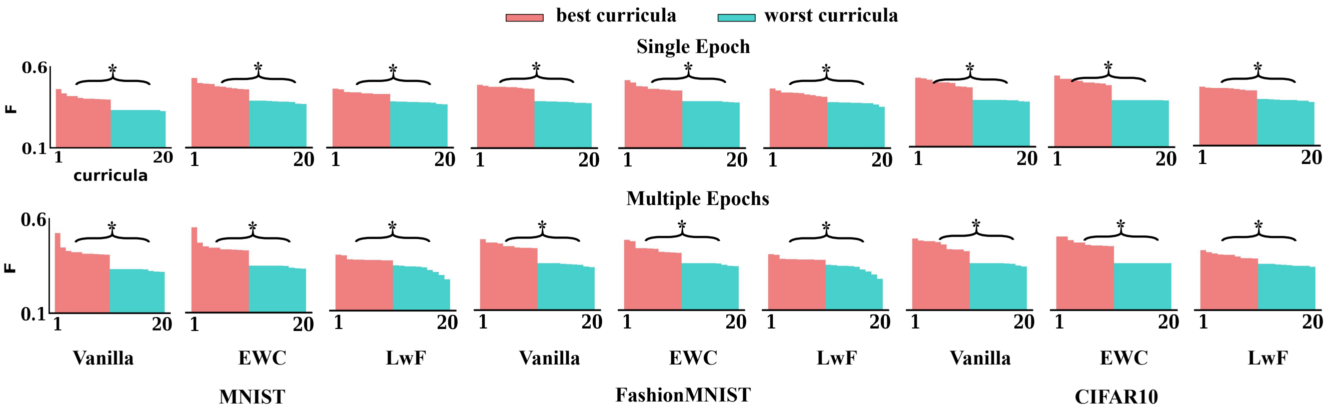


Figure S23: **Top 10 vs bottom 10 curricula across three datasets and three CL algorithms (Sec 3.2) in paradigm-I (5 classes, Sec 3.1).** The top row of plots shows the online setting (single epoch per task), and the bottom row shows the offline setting with multiple epochs per task. Each plot shows the $\mathcal{F}$ scores for the best 10 curricula (red) and the worst 10 curricula (blue). The statistical significance (* = statistically significant) was determined using two-sample t-tests on the 10 best and 10 worst $\mathcal{F}$ scores. See **Sec S4**, and **S11** for results on the impact of curricula in paradigm-I.

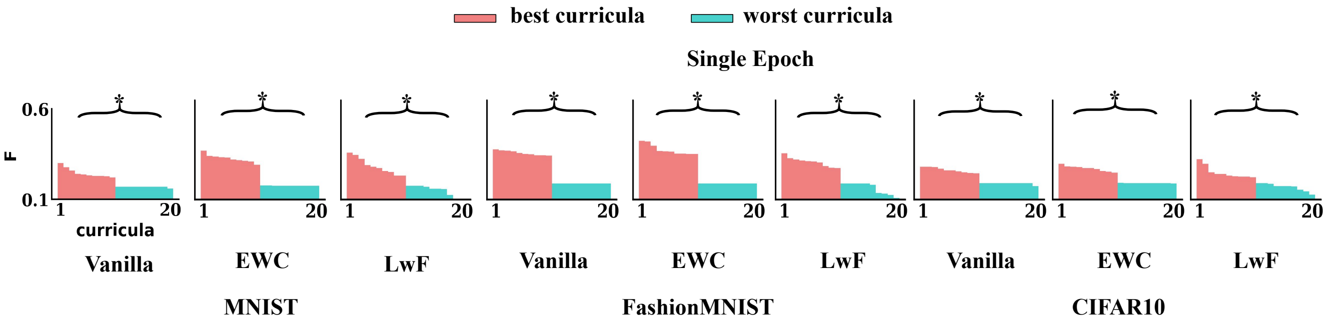


Figure S24: **Top 10 vs bottom 10 curricula across three datasets and three CL algorithms (Sec 3.2) in paradigm-II (10 classes arranged into 5 binary tasks, Sec 3.1).** See **Fig S23** for figure design conventions. See **Sec S4** for results on the impact of curricula in paradigm-II.

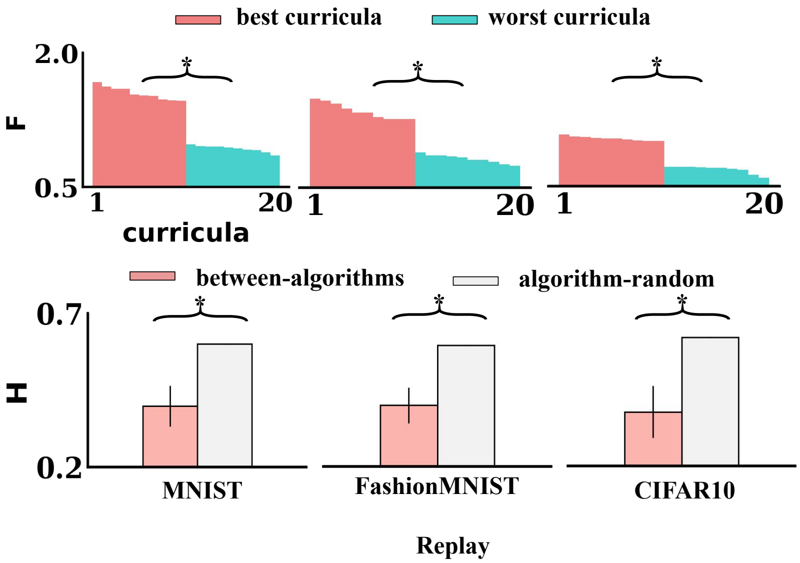


Figure S25: **Top 10 vs bottom 10 curricula, and curriculum discrepancy $\mathcal{H}$ (Sec 3.3), for a naive replay CL algorithm across three datasets in paradigm-I (5 classes, Sec 3.1)**. See **Sec.3.2** for the introduction to the naive replay CL algorithm. Each plot in the top row shows the $\mathcal{F}$ scores for the best 10 curricula (red) and the worst 10 curricula (blue). The second row shows curriculum agreement plots for each dataset (see **Sec S10** for details). Statistical significance (* = statistically significant) was determined using two-sample t-tests between the 10 highest and 10 lowest $\mathcal{F}$ or $\mathcal{H}$ scores (**Sec S13**). The errorbars are the standard deviations over all the test trials. The errorbars are small; and hence, they become almost invisible. See **Sec.S13** for statistical interpretations and analysis.

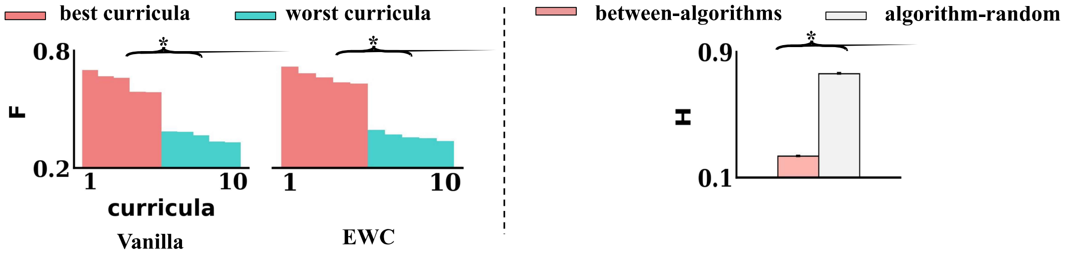


Figure S26: **Strong curriculum effects are observed in the continual visual question answering (VQA) setting.** The left panel shows $\mathcal{F}$ scores for the best 5 and worst 5 curricula using the Vanilla and EWC CL algorithms. and the curricula agreement $\mathcal{H}$ plot for continual VQA (**Sec S12**) on the CLOVER dataset [4]. See **Sec S12** for further analysis and details of VQA experiments.

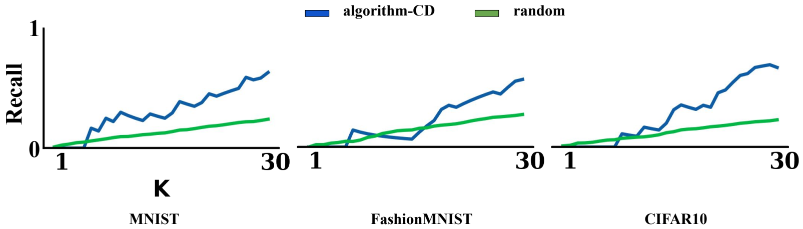
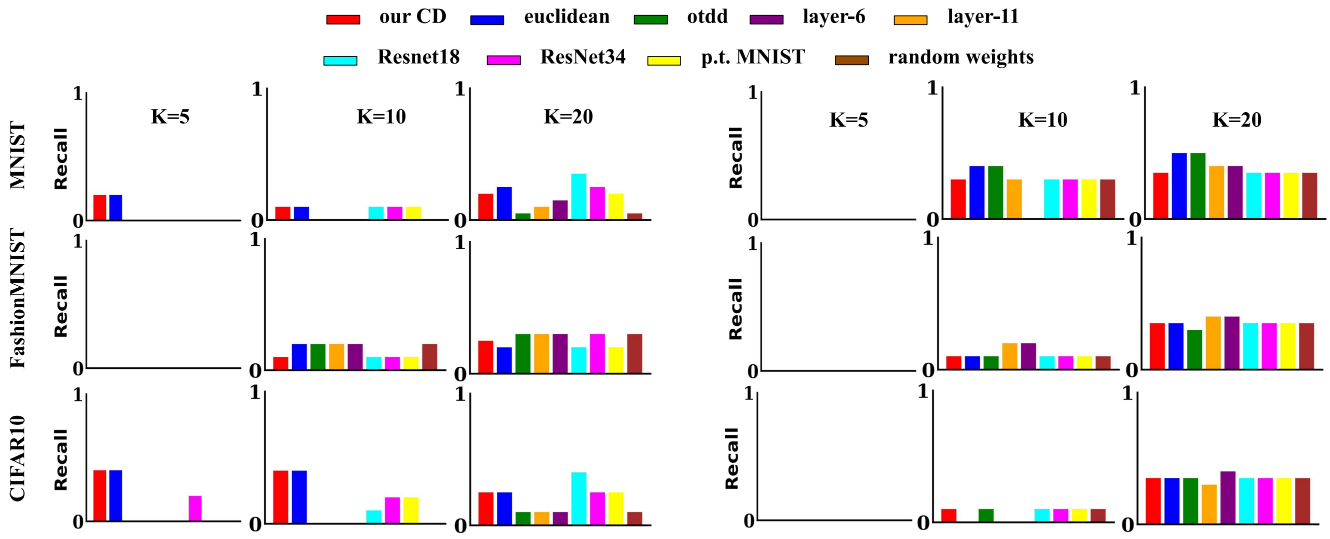


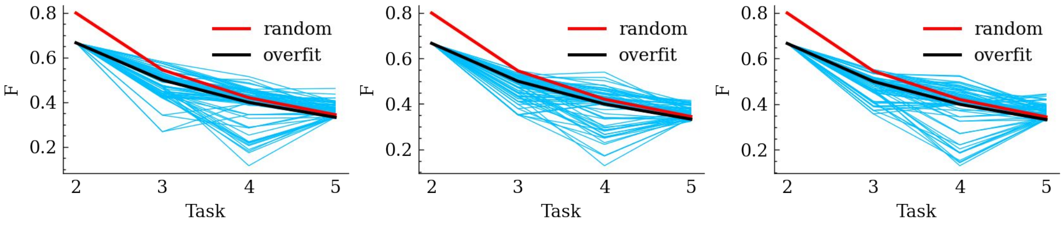
Figure S27: **Our curriculum designer (CD) predicts optimal curricula more accurately than a random CD in paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** See **Fig 4** for the equivalent plots for paradigm-I. The results are analysed in **Sec S7.**



(A) Paradigm-I (5 classes)

(B) Paradigm-II (10 classes)

Figure S28: **Ablation study results on our CD in (A) paradigm-I (5 classes, Sec 3.1) and (B) paradigm-II (10 classes arranged in 5 binary tasks, Sec 3.1).** See **Fig 5** for the same design convention. The results in paradigm-I are analyzed in **Sec 5.3**, and the results in paradigm-II are analyzed in **Sec S9**.



MNIST

FashionMNIST

CIFAR10

Figure S29: **Task-wise $\mathcal{F}$ of the Vanilla CL algorithm across three datasets in paradigm-I (5 classes, Sec 3.1).** Task-wise $\mathcal{F}$ is shown as a blue line plot for each of the $5! = 120$ possible curricula on MNIST, FashionMNIST and CIFAR10. The performance on each task of the random model (red) and a completely over-fitting CL algorithm (black) are also shown (**Sec S5**). See **Sec S5** for the baseline introductions.

Algorithm 1: Python-style pseudocode for CD

```
# N: number of classes
# M (N $\times$ N): M[i][j] is the distance between the feature prototypes of class i and class j;
# Var():  function to compute variance
# C: a given curriculum in sequence of C[1], C[2],...C[i],...,C[N], where i is the class index
# initialize ranking score s
s = 0
# at i = 1
s = 1 - Var(M[1][j]$_{j=2}^{N}$)
for t in (2, N):
    if t $\leq$ ($\lfloor \frac{N}{2} \rfloor$)
        s += M[t][t-1]
    if t > ($\lfloor \frac{N}{2} \rfloor$)
        s += 1 - M[t][N - t + 1]
```

References

[1] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 4

[2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[3] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 3

[4] Stan Weixian Lei, Difei Gao, Jay Zhangjie Wu, Yuxuan Wang, Wei Liu, Mengmi Zhang, and Mike Zheng Shou. Symbolic replay: Scene graph as prompt for continual learning on vqa task. *arXiv preprint arXiv:2208.12037*, 2022. 6, 23

[5] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 3