Full Results Table for the paper "*Protected Attributes Tell Us Who, Behavior Tells Us How: A Comparison of Demographic and Behavioral Oversampling for Fair Student Success Modeling*"

RQ1: Bias Investigation

The following tables show the auc, balanced accuracy, FNR and FPR rate over the entire dataset, and then over different subgroups taken from the entire dataset.
We acknowledge more than 2 genders, and there are more than 2 countries in our Flipped classroom dataset. We use those other categories as groups of their own in our mitigation techniques. Unfortunately, because of the small number of samples belonging in the categories not reported in our table, the metrics are too sensitive to the (mis)classification of a single sample.

TUGLET	overall dataset	School A	School B	difference between areas	females	males	difference between genders
auc	0.68	0.69	0.63	0.06	0.74	0.66	0.08
balanced accuracy	0.61	0.62	0.57	0.05	0.64	0.57	0.07
FNR	0.61	0.57	0.73	0.16	0.53	0.7	0.17
FPR	0.16	0.2	0.14	0.06	0.18	0.17	0.01

Table 1. Baseline results for TugLet. Average ROC, balanced accuracy, false negative rates and false positive rates across the 10 cross validation folds for computed first on the entire student population, than for the students in rancho, the students in tierra linda, the female students and the male students separately. The difference between area is the difference in results between people from rancho and tierra linda. The difference between gender is the difference.

	overall dataset	females	males	difference between genders	Country 1	Country 2	difference between country of diploma
auc	0.63	0.62	0.67	0.05	0.58	0.7	0.12
balanced accuracy	0.59	0.57	0.62	0.05	0.55	0.6	0.05
FNR	0.46	0.56	0.43	0.13	0.58	0.42	0.16
FPR	0.35	0.35	0.33	0.02	0.33	0.38	0.06

Table 2. Baseline results for Flipped Classroom. Average ROC, balanced accuracy, false negative rates and false negative rates across the 10 cross validation folds for computed first on the entire student population, then for the female students, the male students, the students who received their diploma in France and those who received it from Switzerland separately. The difference between genders is the difference in results between females and males. The difference between country of diploma is the difference across the students who received their diploma from France and those who received it from Switzerland.

RQ2: Attribute Oversampling

Those tables report the mean auc, balanced accuracy, FNR and FPR over a stratified 10-fold cross validation. For each experiment, the bolded attributes are those used to oversample the data

	overall	School A	School B	difference between areas	females	males	differences between gender	mode
auc: **baseline**	0.68	0.69	0.63	0.06	0.74	0.66	0.08	none
auc: **area** & **label**	0.67	0.68	0.64	0.04	0.7	0.65	0.05	minor
auc: **gender** & **label**	0.68	0.67	0.65	0.02	0.74	0.65	0.09	cascade
auc: **area** & **gender** & **label**	0.67	0.68	0.61	0.07	0.68	0.66	0.02	equal
balanced accuracy: **baseline**	0.61	0.62	0.57	0.05	0.64	0.57	0.07	none
balanced accuracy: **area** & **label**	0.62	0.6	0.64	0.04	0.63	0.61	0.02	minor
balanced accuracy: **gender** & **label**	0.61	0.6	0.58	0.02	0.65	0.58	0.07	cascade
balanced accuracy: **area** & **gender** & **label**	0.62	0.6	0.6	0	0.64	0.6	0.04	equal
false negative: **baseline**	0.61	0.57	0.73	0.16	0.53	0.7	0.17	none
false negative: **area** & **label**	0.39	0.41	0.41	0	0.32	0.46	0.14	minor
false negative: **gender** & **label**	0.64	0.63	0.72	0.09	0.56	0.71	0.15	cascade
false negative: **area** & **gender** & **label**	0.44	0.45	0.5	0.05	0.4	0.48	0.08	equal
false positive: **baseline**	0.16	0.2	0.14	0.06	0.18	0.17	0.01	none
false positive: **area** & **label**	0.36	0.38	0.32	0.06	0.42	0.32	0.1	minor
false positive: **gender** & **label**	0.15	0.18	0.12	0.06	0.15	0.13	0.02	cascade
false positive: **area** & **gender** & **label**	0.32	0.35	0.3	0.05	0.33	0.32	0.01	equal

Table 3. Demographics oversampling results for TugLet. Each row represents a new experiment, where the bolded term are the attribute used to compute the upsampling ratios, with the exception of baseline where no oversampling was conducted.

FLIPPED	overall dataset	females	males	difference between genders	Country 1	Country 2	difference between country of diploma	mode
auc: baseline	0.63	0.62	0.67	0.05	0.58	0.7	0.12	
auc: gender	0.64	0.67	0.63	0.04	0.61	0.68	0.07	major
auc: intervention	0.7	0.68	0.72	0.04	0.69	0.75	0.06	major
auc: country & gender	0.67	0.75	0.63	0.12	0.64	0.69	0.05	equal
auc: country & intervention	0.64	0.64	0.64	0	0.51	0.73	0.22	major
balanced accuracy: baseline	0.59	0.57	0.62	0.05	0.55	0.6	0.05	
balanced accuracy: gender	0.61	0.61	0.61	0	0.51	0.63	0.12	major
balanced accuracy: intervention	0.64	0.61	0.65	0.04	0.63	0.68	0.05	major
balanced accuracy: country & gender	0.62	0.67	0.6	0.07	0.61	0.64	0.03	equal
balanced accuracy: country & intervention	0.61	0.6	0.62	0.02	0.58	0.66	0.08	major
FNR: baseline	0.46	0.56	0.43	0.13	0.58	0.42	0.16	
FNR: gender	0.43	0.49	0.42	0.07	0.61	0.35	0.26	major
FNR: intervention	0.45	0.56	0.43	0.13	0.5	0.42	0.08	major
FNR: country & gender	0.45	0.47	0.45	0.02	0.47	0.45	0.02	equal
FNR: country & intervention	0.41	0.52	0.38	0.14	0.49	0.34	0.15	major
FPR: baseline	0.35	0.35	0.33	0.02	0.33	0.38	0.05	
FPR: gender	0.36	0.32	0.37	0.05	0.36	0.38	0.02	major
FPR: intervention	0.25	0.25	0.26	0.01	0.23	0.23	0	major
FPR: country & gender	0.31	0.23	0.35	0.12	0.31	0.28	0.03	equal
FPR: country & intervention	0.36	0.31	0.39	0.08	0.35	0.35	0	major

Table 4. Demographics oversampling results for Flipped Classroom. Each row represents a new experiment, where the bolded term are the attribute used to compute the upsampling ratios, with the exception of baseline where no oversampling was conducted.

RQ3: Behavioral Oversampling

Those tables report the mean auc, balanced accuracy, FNR and FPR over a stratified 10-fold cross validation. For each experiment, the bolded attributes denotes whether the clustering labels were used alone or along with other attributes.

	overall	**School A**	**School B**	**difference between areas**	**females**	**males**	**differences between gender**	**mode**
auc: **baseline**	0.68	0.69	0.63	0.06	0.74	0.66	0.08	none
auc: **cluster**	0.67	0.65	0.6	0.05	0.72	0.63	0.09	minor
auc: **cluster** & **label**	0.66	0.64	0.63	0.01	0.66	0.65	0.01	equal
auc: **cluster** & **gender** & **label**	0.67	0.68	0.61	0.07	0.71	0.64	0.07	within
balanced accuracy: **baseline**	0.61	0.62	0.57	0.05	0.64	0.57	0.57	none
balanced accuracy: **cluster**	0.6	0.6	0.56	0.04	0.63	0.57	0.06	minor
balanced accuracy: **cluster** & **label**	0.57	0.58	0.53	0.05	0.55	0.61	0.06	equal
balanced accuracy: **cluster** & **gender** & **label**	0.63	0.63	0.58	0.05	0.62	0.63	0.01	within
false negative: **baseline**	0.61	0.57	0.73	0.16	0.53	0.7	0.17	none
false negative: **cluster**	0.55	0.54	0.64	0.1	0.5	0.59	0.09	minor
false negative: **cluster** & **label**	0.5	0.46	0.6	0.14	0.5	0.5	0	equal
false negative: **cluster** & **gender** & **label**	0.42	0.38	0.56	0.18	0.4	0.45	0.05	within
false positive: **baseline**	0.16	0.2	0.14	0.06	0.18	0.17	0.01	none
false positive: **cluster**	0.25	0.26	0.24	0.02	0.25	0.27	0.02	minor
false positive: **cluster** & **label**	0.36	0.37	0.34	0.03	0.41	0.29	0.11	equal
false positive: **cluster** & **gender** & **label**	0.31	0.37	0.28	0.09	0.35	0.3	0.05	within

Table 5. Cluster oversampling results for TugLet. Average ROC, balanced accuracy, false negative rates and false negative rates across the 10 cross validation folds. Each row represents a new experiment, where the bolded term are the attribute used to compute the upsampling ratios, with the exception of baseline where no oversampling was conducted.

	overall dataset	**females**	**males**	**difference between genders**	**Country 1**	**Country 2**	**difference between country of diploma**	**mode**
auc: baseline	0.63	0.62	0.67	0.05	0.58	0.7	0.12	-
auc: gender	0.64	0.67	0.63	0.04	0.61	0.68	0.07	major
auc: intervention	0.7	0.68	0.72	0.04	0.69	0.75	0.06	major
auc: cluster	0.66	0.69	0.67	0.02	0.6	0.7	0.1	cascade
auc: country & intervention	0.64	0.64	0.64	0	0.51	0.73	0.22	major
auc: cluster & intervention	0.7	0.75	0.67	0.08	0.65	0.73	0.08	equal balancing
balanced accuracy: baseline	0.59	0.57	0.62	0.05	0.55	0.6	0.05	-
balanced accuracy: gender	0.61	0.61	0.61	0	0.51	0.63	0.12	major
balanced accuracy: intervention	0.64	0.61	0.65	0.04	0.63	0.68	0.05	major
balanced accuracy: cluster	0.63	0.69	0.6	0.09	0.61	0.66	0.05	cascade
balanced accuracy: country & intervention	0.61	0.6	0.62	0.02	0.58	0.66	0.08	major
balanced accuracy: cluster & intervention	0.67	0.69	0.66	0.03	0.63	0.67	0.04	equal balancing
FNR: baseline	0.46	0.56	0.43	0.13	0.58	0.42	0.16	-
FNR: gender	0.43	0.49	0.42	0.07	0.61	0.35	0.26	major
FNR: intervention	0.45	0.56	0.43	0.13	0.5	0.42	0.08	major
FNR: cluster	0.43	0.44	0.43	0.01	0.47	0.38	0.09	cascade
FNR: country & intervention	0.41	0.52	0.38	0.14	0.49	0.34	0.15	major
FNR: cluster & intervention	0.39	0.48	0.35	0.13	0.46	0.34	0.12	equal balancing
FPR: baseline	0.35	0.35	0.33	0.02	0.33	0.38	0.05	-
FPR: gender	0.36	0.32	0.37	0.05	0.36	0.38	0.02	major
FPR: intervention	0.25	0.25	0.26	0.01	0.23	0.23	0	major
FPR: cluster	0.32	0.22	0.37	0.15	0.3	0.29	0.01	cascade
FPR: country & intervention	0.36	0.31	0.39	0.08	0.35	0.35	0	major
FPR: cluster & intervention	0.28	0.17	0.33	0.16	0.28	0.32	0.04	equal balancing

Table 6. Cluster oversampling results for Flipped Classroom. Average ROC, balanced accuracy, false negative rates and false negative rates across the 10 cross validation folds. Each row represents a new experiment, where the bolded term are the attribute used to compute the upsampling ratios, with the exception of baseline where no oversampling was conducted.