Completely Blind Quality Assessment of User Generated Video Content

Invited Talk at SPCOM 2022

Sumohana S. Channappayya

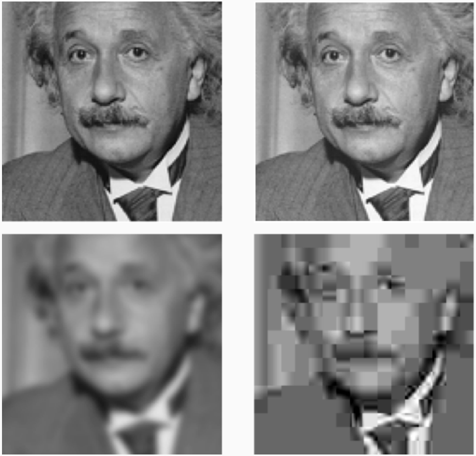
July 13, 2022

Indian Institute of Technology Hyderabad

Outline

- Introduction to Quality Assessment
- Perceptual Straightening Hypothesis
- Proposed Blind Video Quality Assessment Algorithm
- Results
- Concluding Remarks

What is Quality Assessment?



Why Quality Assessment?

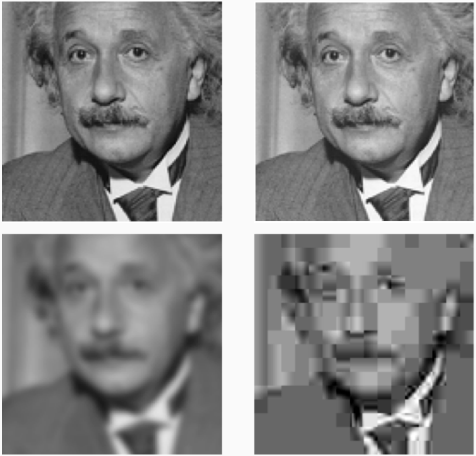


Table 1: Distorted images have same mean squared error (MSE)! L^p norms fail! [12] [1]

[1] Z Wang and A C Bovik. “Mean squared error: Love it or leave it? A new look at signal fidelity measures”. In: *IEEE Signal Processing Magazine* 26.1 (2009), pp. 98–117.

Why Quality Assessment?

Figure 1: Guarantee visual quality. Why? $\approx$ 6.6 billion smart-phones in 2021![2]

[2] http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[3] http://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html

Why Quality Assessment?

Figure 1: Guarantee visual quality. Why? $\approx$ 6.6 billion smart-phones in 2021![2]

Figure 2: Optimal resource usage. Why? More than 1 trillion photos per year in recent years![3]

[2] http://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

[3] http://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html

Background

Human Perceptual Quality

Measured in terms of *mean opinion score (MOS)* and *difference MOS (DMOS)*. It forms the *ground truth* in all QA work.

Background

Human Perceptual Quality

Measured in terms of *mean opinion score (MOS)* and *difference MOS (DMOS)*. It forms the *ground truth* in all QA work.

Multimedia Quality Assessment

An *objective method* to estimate *human perceptual quality* or *MOS* of test content M_{test} either in the presence of pristine content M_{ref}, its estimate $\hat{M}_{ref}$ or in a stand-alone mode

Background

Human Perceptual Quality

Measured in terms of *mean opinion score (MOS)* and *difference MOS (DMOS)*. It forms the *ground truth* in all QA work.

Multimedia Quality Assessment

An *objective method* to estimate *human perceptual quality* or *MOS* of test content M_{test} either in the presence of pristine content M_{ref}, its estimate $\hat{M}_{ref}$ or in a stand-alone mode

- Full-reference (FR): $Q_{test} = f(M_{ref}, M_{test}; \theta)$

Background

Human Perceptual Quality

Measured in terms of *mean opinion score (MOS)* and *difference MOS (DMOS)*. It forms the *ground truth* in all QA work.

Multimedia Quality Assessment

An *objective method* to estimate *human perceptual quality* or *MOS* of test content M_{test} either in the presence of pristine content M_{ref}, its estimate $\hat{M}_{ref}$ or in a stand-alone mode

- Full-reference (FR): $Q_{test} = f(M_{ref}, M_{test}; \theta)$
- Reduced-reference (RR): $Q_{test} = g(\hat{M}_{ref}, M_{test}; \theta)$

Background

Human Perceptual Quality

Measured in terms of *mean opinion score (MOS)* and *difference MOS (DMOS)*. It forms the *ground truth* in all QA work.

Multimedia Quality Assessment

An *objective method* to estimate *human perceptual quality* or *MOS* of test content M_{test} either in the presence of pristine content M_{ref}, its estimate $\hat{M}_{ref}$ or in a stand-alone mode

- Full-reference (FR): $Q_{test} = f(M_{ref}, M_{test}; \theta)$
- Reduced-reference (RR): $Q_{test} = g(\hat{M}_{ref}, M_{test}; \theta)$
- **No-reference (NR):** $Q_{test} = h(M_{test}; \theta)$

Performance Measures

- Linear Correlation Coefficient (LCC)
- Spearman Rank Ordered Correlation Coefficient (SROCC)
- Root Mean Squared Error (RMSE)

Blind Video Quality Assessment

- How do we assess the perceptual quality of natural videos in the blind (NR) setting?
- **Natural videos have rich temporal information**
- How do we leverage this rich temporal information?
- **Straightness principle:** Predictions of future samples can be formulated as linear operations in the latent space/representation space[4]

[4] Goroshin, Ross, Mathieu, Michael, and LeCun, Yann. "Learning to linearize under uncertainty." NeurIPS 2015.

Perceptual Straightening in the Human Visual System

Perceptual Straightening Hypothesis [2] [5]

"Many behaviors rely on predictions derived from recent visual input, but the temporal evolution of those inputs is generally complex and difficult to extrapolate. We propose that the visual system transforms these inputs to follow straighter temporal trajectories."

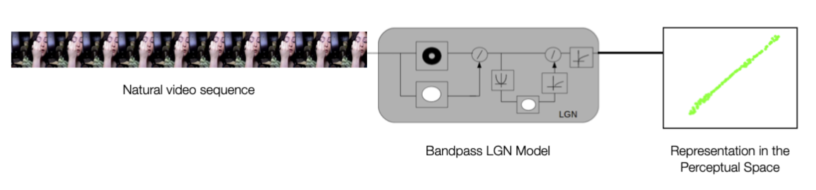


Figure 3: Illustration of the perceptual straightening hypothesis

[5] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. "Perceptual straightening of natural videos". In: *Nature Neuroscience* 22.6 (2019), p. 984.

Two Simple Questions

1. What happens to the perceptual straightness of distorted natural videos?
2. Is the perceptual straightness a function of video quality?

Empirical Analysis of Q1

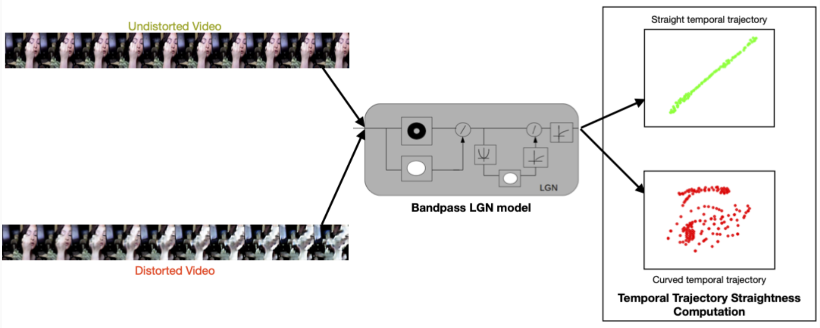


Figure 4: Effect of distortion on perceptual domain representation

Empirical Analysis of Q2

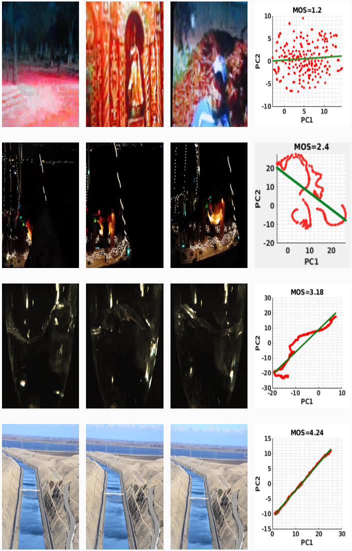


Figure 5: Straightness increases with MOS

Completely Blind Video Quality Assessment

Temporal Quality Estimation

- The input frame X_t is passed through the LGN model to find the perceptual domain representation F_t, followed by PCA-based dimensionality reduction to give a d-dimensional vector $\mathbf{f}_t$
- Estimate the feature at current time using a linear model

$$\hat{\mathbf{f}}_t = \beta_0 \mathbf{1} + \sum_{i=1}^{K} \beta_i \mathbf{f}_{t-i},$$

  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$: scalar model parameters
  - $\mathbf{1}$: d-dimensional vector of ones
  - $\mathbf{f}_t$: ground truth representation of frame at time t
  - $\hat{\mathbf{f}}_t$: prediction at time t
  - K is a tunable parameter

Completely Blind Video Quality Assessment

Temporal Quality Estimation

- Frame level error:

$$\mathrm{D}_t = ||\mathbf{f}_t - \hat{\mathbf{f}}_t||_1$$

- Temporal quality estimate over an N-frame video:

$$Q_{\mathrm{temporal}} = \log\left(\frac{1}{N}\sum_{t=1}^{N} \mathrm{D}_t\right)$$

Completely Blind Video Quality Assessment

Spatial Quality Estimation [6]

- Estimate the frame level quality

$$q_t = \sum_{i=t-\frac{N}{3}+1}^{t} w_{t-i+1} \cdot \mathrm{NIQE}(X_i),$$

where $\mathrm{NIQE}(X_i)$ is the NIQE score of frame X_i,

$$w_j = \frac{\exp(-\alpha j)}{\sum_{i=1}^{\frac{N}{3}} \exp(-\alpha i)}, 1 \leq j \leq \frac{N}{3}.$$

- Spatial quality estimate over an N-frame video:

$$Q_{\mathrm{spatial}} = \frac{1}{N} \sum_{i=1}^{N} q_i.$$

[6] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C.Bovik, "A comparative evaluation of temporal pooling methods for blind video qua- Spatial Quality" arXiv preprint arXiv:2002.10651, 2020

STraightness Evaluation Metric (STEM)

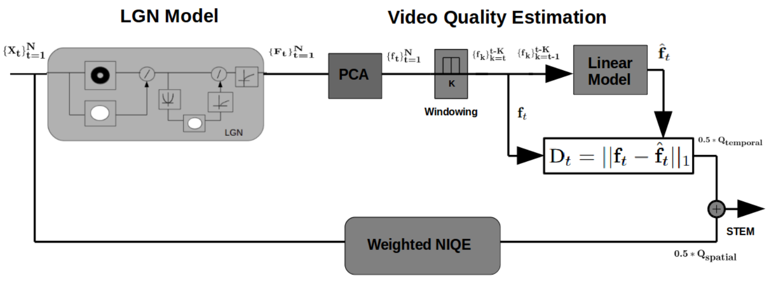


Figure 6: Block diagram of the proposed BVQA algorithm STEM.

$$\text{STEM} = \frac{Q_{\text{temporal}} + Q_{\text{spatial}}}{2}$$

User Generated Video Quality Assessment Datasets

- **KoNViD-1K dataset [3]:** 1200 videos, $> 960 \times 540$, 8 sec, 24/25/30 fps, various attributes (blur, contrast, colourfulness, etc.), CC attributed source videos, Crowdsourced, MOS
- **VQC dataset [9]:** 585 videos, 404×720, 10 sec, HD, Full HD, 43 mobile devices, Crowdsourced (AMT), MOS
- **CVD dataset [7]:** 234 videos, QCIF to Full HD, 10-25 sec, 10-31 fps, 78 cameras, Crowdsourced, MOS
- **YouTube-UGC dataset [11]:** 1380 videos, 360p to 4K, 20 sec, Gaming, Sports, Music Video etc., Crowdsourced, MOS
- **LIVE Qualcomm dataset [1]:** 208 videos, Full HD, 8 cameras, 15 sec, 30 fps, Crowdsourced, MOS, artifacts, color, exposure, focus, sharpness, stabilization

Performance Evaluation on the KoNViD-1K Dataset [3]

Table 2: Performance evaluation results and comparison with representative supervised (italics) and unsupervised/completely blind VQA algorithms on the KoNViD-1K dataset [3] ($K = 6$, $d = 10$)

Method	LCC	SROCC	RMSE
V-BLIINDS [8]	*0.565*	*0.572*	*0.526*
TL-VQM [4]	*0.79*	*0.80*	*0.406*
VIIDEO [5]	-0.015	0.013	0.639
NIQE [6]	0.544	0.542	0.537
NIQE Hysteresis pooling [10]	0.563	0.569	-
Q_{temporal}	0.444	0.450	0.574
Q_{spatial}	0.547	0.546	0.534
STEM	**0.629**	**0.629**	**0.497**

Performance on the VQC Dataset [9]

Table 3: Performance evaluation results and comparison with representative supervised (italics) and unsupervised/completely blind VQA algorithms on the LIVE VQC dataset [9] ($K = 6$, $d = 10$).

Method	LCC	SROCC	RMSE
V-BLIINDS [8]	*0.718*	*0.707*	*11.546*
VIIDEO [5]	0.137	0.029	16.882
NIQE [6]	0.610	0.563	13.890
NIQE Percentile pooling [10]	0.630	0.634	-
Q_{temporal}	0.454	0.466	15.148
Q_{spatial}	0.613	0.594	13.467
STEM	**0.670**	**0.656**	**12.649**

Performance Evaluation on the CVD Dataset [7]

Table 4: Performance evaluation results and comparison with representative supervised (italics) and unsupervised/completely blind VQA algorithms on the CVD dataset [7] ($K = 6$, $d = 10$).

Method	LCC	SROCC	RMSE
V-BLIINDS [8]	*0.71*	*0.70*	*15.222*
TL-VQM [4]	*0.85*	*0.83*	*11.33*
VIIDEO [5]	-	-	
NIQE [6]	0.61	0.58	17.15
Q_{temporal}	0.361	0.355	20.507
Q_{spatial}	0.619	0.580	16.834
STEM	**0.629**	**0.593**	**16.664**

Performance Evaluation on the YouTube-UGC Dataset [11]

Table 5: Performance evaluations results and comparison with supervised (italics) and unsupervised/completely blind VQA algorithms on the YouTube-UGC dataset [11] ($K = 6$, $d = 10$).

Method	LCC	SROCC	RMSE
V-BLIINDS [8]	0.559	0.555	0.535
VIIDEO [5]	0.146	0.130	0.637
NIQE [6]	0.105	0.236	0.640
Q_{temporal}	0.272	0.321	0.636
Q_{spatial}	0.286	0.239	0.6221
STEM	**0.318**	**0.284**	**0.623**

Performance Evaluation on the LIVE Qualcomm Dataset [1]

Table 6: LCC performance results on the LIVE Qualcomm dataset [1] ($K = 6, d = 10$). Representative supervised VQA algorithms are in italics.

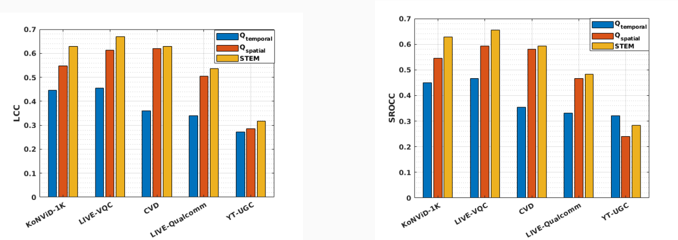
Method	artifacts	color	focus	sharpness	stabilization	exposure	all
V-BLIINDS [8]	*0.8386*	*0.664*	*0.807*	*0.684*	*0.713*	*0.690*	*0.665*
TL-VQM [4]	-	-	-	-	-	-	*0.81*
VIIDEO [5]	0.288	0.331	0.251	0.3012	0.369	0.207	0.098
Q_{temporal}	0.566	0.304	0.280	0.475	0.423	0.749	0.339
Q_{spatial}	0.4638	0.4703	0.4523	0.619	0.596	0.526	0.504
STEM	**0.725**	**0.493**	**0.563**	**0.638**	**0.631**	**0.587**	**0.537**

Performance Evaluation on the LIVE Qualcomm Dataset [1]

Table 7: SROCC performance results on the LIVE Qualcomm dataset [1] ($K = 6, d = 10$). Representative supervised VQA algorithms are in italics.

Method	artifacts	color	focus	sharpness	stabilization	exposure	all
V-BLIINDS [8]	*0.732*	*0.607*	*0.803*	*0.678*	*0.660*	*0.642*	*0.617*
TL-VQM [4]	-	-	-	-	-	-	0.84
VIIDEO [5]	-0.178	0.142	0	-0.178	-0.107	-0.071	-0.141
Q_{temporal}	0.432	0.381	0.428	0.491	0.521	0.446	0.332
Q_{spatial}	0.450	0.341	0.556	0.505	0.338	0.297	0.467
STEM	**0.646**	**0.527**	**0.549**	**0.593**	**0.412**	**0.555**	**0.483**

The Contributions of Q_{temporal} and Q_{spatial} to Performance



(a) LCC values for the UGC-datasets

(b) SROCC values for the UGC-datasets

Figure 7: Bar graphs illustrating the ablation study involving the components Q_{temporal}, Q_{spatial} and their combination in STEM on the five UGC datasets considered in this work.

Concluding Remarks

- Explainable approach to NRVQA - inspired by the idea of perceptual straightening
- STEM is completely blind and it is computationally not very expensive
- Few parameters in the computational models
- STEM delivers competitive performance on the authentic VQA datasets

Acknowledgements

- Part of Parimala Kancharla's PhD work
- Thanks to the SPCOM 2022 organizers for the invitation!

References

[1] Deepti Ghadiyaram et al. "In-capture mobile video distortions: A study of subjective behavior and objective algorithms". In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.9 (2017), pp. 2061–2077.

[2] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. "Perceptual straightening of natural videos". In: *Nature neuroscience* 22.6 (2019), p. 984.

[3] Vlad Hosu et al. "The Konstanz natural video database (KoNViD-1k)". In: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2017, pp. 1–6.

[4] Jari Korhonen. "Two-level approach for no-reference consumer video quality assessment". In: *IEEE Transactions on Image Processing* 28.12 (2019), pp. 5923–5938.

[5] Anish Mittal, Michele A Saad, and Alan C Bovik. "A completely blind video integrity oracle". In: *IEEE Transactions on Image Processing* 25.1 (2016), pp. 289–300.

[6] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. “Making a “completely blind” image quality analyzer”. In: *IEEE Signal Processing Letters* 20.3 (2012), pp. 209–212.

[7] Mikko Nuutinen et al. “CVD2014—A database for evaluating no-reference video quality assessment algorithms”. In: *IEEE Transactions on Image Processing* 25.7 (2016), pp. 3073–3086.

[8] Michele A Saad, Alan C Bovik, and Christophe Charrier. “Blind prediction of natural video quality”. In: *IEEE Transactions on Image Processing* 23.3 (2014), pp. 1352–1365.

[9] Zeina Sinno and Alan Conrad Bovik. “Large-scale study of perceptual video quality”. In: *IEEE Transactions on Image Processing* 28.2 (2018), pp. 612–627.

[10] Zhengzhong Tu et al. “A Comparative Evaluation Of Temporal Pooling Methods For Blind Video Quality Assessment”. In: *2020 IEEE International Conference on Image Processing (ICIP)*. 2020, pp. 141–145. DOI: *10.1109/ICIP40778.2020.9191169*.

[11] Yilin Wang, Sasi Inguva, and Balu Adsumilli. “YouTube UGC dataset for video compression research”. In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2019, pp. 1–5.

[12] Zhou Wang and Alan C Bovik. “Mean squared error: Love it or leave it? A new look at signal fidelity measures”. In: *IEEE signal processing magazine* 26.1 (2009), pp. 98–117.

Tuning of the Parameters K, d

- Performance of the proposed blind VQA algorithm on various UGC datasets different values of the hyperparameter K with $d = 10$

Dataset	**Method**	$K=2$			$K=4$			$K=6$			$K=8$		
		LCC	SROCC	RMSE	LCC	SROCC	RMSE	LCC	SROCC	RMSE	LCC	SROCC	RMSE
KoNViD-1K	Q_{temporal}	0.426	0.435	0.579	0.441	0.448	0.572	0.444	0.450	0.574	0.446	0.451	0.574
	STEM	**0.628**	**0.628**	**0.498**	**0.635**	**0.633**	**0.498**	**0.629**	**0.629**	**0.497**	**0.628**	**0.626**	**0.498**
LIVE-VQC	Q_{temporal}	0.460	0.462	15.143	0.461	0.467	14.553	0.459	0.466	15.148	0.448	0.457	15.143
	STEM	**0.671**	**0.657**	**12.639**	**0.669**	**0.655**	**12.399**	**0.670**	**0.656**	**12.649**	**0.670**	**0.653**	**12.639**
CVD	Q_{temporal}	0.29	0.279	20.519	0.355	0.350	20.037	0.361	0.315	20.507	0.264	0.263	20.651
	STEM	**0.626**	**0.588**	**16.717**	**0.623**	**0.593**	**16.905**	**0.629**	**0.593**	**16.664**	**0.619**	**0.567**	**16.834**
LIVE-Qualcomm	Q_{temporal}	0.275	0.241	11.745	0.339	0.332	11.702	0.285	0.244	11.745	0.299	0.281	11.901
	STEM	**0.537**	**0.483**	**10.125**	**0.537**	**0.483**	**10.102**	**0.537**	**0.483**	**10.102**	**0.537**	**0.483**	**10.101**
YouTube - UGC	Q_{temporal}	0.218	0.213	0.631	0.225	0.215	0.636	0.272	0.321	0.6273	0.269	0.327	0.628
	STEM	**0.295**	**0.294**	**0.624**	**0.296**	**0.292**	**0.623**	**0.318**	**0.284**	**0.618**	**0.318**	**0.285**	**0.618**

- Performance of the proposed blind VQA algorithm on various UGC datasets different values of the hyperparameter d with $K = 6$

Dataset	**Method**	$d = 10$			$d = 30$			$d = 50$			$d = 80$		
		LCC	SROCC	RMSE	LCC	SROCC	RMSE	LCC	SROCC	RMSE	LCC	SROCC	RMSE
KoNViD-1K	Q_{temporal}	0.444	0.450	0.574	0.440	0.446	0.575	0.440	0.451	0.574	0.436	0.435	0.575
	STEM	**0.629**	**0.629**	**0.497**	**0.627**	**0.626**	**0.499**	**0.629**	**0.629**	**0.497**	**0.628**	**0.628**	**0.498**
LIVE-VQC	Q_{temporal}	0.459	0.466	15.148	0.455	0.461	15.189	0.450	0.449	15.227	0.459	0.445	15.150
	STEM	**0.670**	**0.656**	**12.649**	**0.665**	**0.648**	**12.726**	**0.664**	**0.645**	**12.753**	**0.663**	**0.649**	**12.765**
CVD	Q_{temporal}	0.361	0.315	20.507	0.324	0.275	20.282	0.332	0.326	20.217	0.345	0.344	20.123
	STEM	**0.629**	**0.593**	**16.664**	**0.626**	**0.564**	**16.928**	**0.614**	**0.562**	**16.919**	**0.662**	**0.582**	**16.721**
LIVE-Qualcomm	Q_{temporal}	0.285	0.244	11.745	0.312	0.254	11.683	0.327	0.324	11.674	0.294	0.255	11.702
	STEM	**0.537**	**0.483**	**10.102**	**0.537**	**0.483**	**10.123**	**0.537**	**0.483**	**10.125**	**0.537**	**0.483**	**10.118**
YouTube - UGC	Q_{temporal}	0.272	0.321	0.627	0.214	0.248	0.646	0.235	0.271	0.646	0.244	0.3516	0.632
	STEM	**0.318**	**0.284**	**0.618**	**0.305**	**0.292**	**0.621**	**0.307**	**0.274**	**0.620**	**0.317**	**0.296**	**0.618**