Journal of Transfer and Generalization
Volume 1 · Number 1 · 2026

Equal Image Counts Do Not Ensure Transferable Reconstruction: A Linear Study of Two Handwritten-Digit Resources

Charles Wallace[1] · Nicholas Foster[2] · Paul Lewis[3]

[1] Lithuania · [2] Hungary · [3] Luxembourg

Original publication: November 1, 2026 | Article 613

Abstract

A common eight-by-eight grid can hide substantial differences between image collections. We fitted two rank-sixteen reconstruction bases, each using 2,000 training examples, and exchanged them between MNIST and Optdigits. Under a fixed central four-cell mask, the transferred basis increased mean squared error by 0.0400 on 10,000 MNIST test images and by 0.1587 on 1,797 Optdigits test images. These are conditional paired comparisons of linear reconstruction policies. Their asymmetry motivates checking the measurement process behind a shared representation before treating equal training counts as evidence of portability.

Research draft. Exploratory public-data analysis; scientific interpretation and publication decisions await accountable author review.

Equal Image Counts Do Not Ensure Transferable Reconstruction: A Linear Study of Two Handwritten-Digit Resources

> Research draft. Exploratory public-data analysis; scientific interpretation and publication decisions await accountable author review.

Abstract

A common eight-by-eight grid can hide substantial differences between image collections. We fitted two rank-sixteen reconstruction bases, each using 2,000 training examples, and exchanged them between MNIST and Optdigits. Under a fixed central four-cell mask, the transferred basis increased mean squared error by 0.0400 on 10,000 MNIST test images and by 0.1587 on 1,797 Optdigits test images. These are conditional paired comparisons of linear reconstruction policies. Their asymmetry motivates checking the measurement process behind a shared representation before treating equal training counts as evidence of portability.

1. Why matching dimensions is an incomplete transfer argument

Restoration projects often begin with a convenient public image collection and later move toward a specialized archive. Standardizing the array dimensions makes the software reusable, but does not establish that corresponding numerical cells represent the same measurement process. One source may store averaged grayscale intensity while another counts foreground pixels after binarization and normalization. A fitted covariance basis absorbs these choices alongside differences in handwriting. Reusing that basis therefore transfers a representation of the acquisition pipeline as well as a representation of digit shape.

The present research question is specific: at an equal fitting count, does a linear basis learned from one public digit resource reconstruct a hidden central region as well as a basis learned locally? MNIST is linked to document-recognition research by Lecun et al., 1998, whereas the Optdigits repository describes a separate block-count construction and writer partition Alpaydin and Kaynak, 1998. Both support numerical experiments without collecting new participants. Neither is an oracle-bone rubbing collection. Their role here is to make an otherwise vague portability claim testable using independently released source partitions.

This question matters for interpreting specialized restoration work. Meng et al., 2025 addresses oracle-bone inscription inpainting using a visual Mamba component; its domain specificity motivates asking which parts of a restoration representation can travel between collections. We do not execute that architecture or estimate its transfer performance. The current experiment deliberately uses a transparent linear reconstruction rule so that a change in the fitted basis can be separated from a change in optimization machinery. It provides a small diagnostic of transfer assumptions, rather than a competing restoration system.

Two analytical propositions organize the study. First, equal sample counts and equal array support need not yield interchangeable reconstruction bases. Second, the size of a transfer penalty may differ by target collection because the two source-to-target directions condition on different observed images. A positive penalty supports a local advantage for the declared procedure. It does not identify which source characteristic produced that advantage. Matching sample counts addresses one design margin; the broader

matching literature makes clear why balance on a selected margin does not establish causal exchangeability Stuart, 2010.

2. What is being transferred

The reconstruction object has two fitted parts: a mean image and sixteen orthonormal covariance directions. The principal-component construction follows the low-dimensional approximation tradition associated with Pearson, 1901. Probabilistic principal-component analysis supplies a related latent-variable interpretation Tipping and Bishop, 1999, although our implementation uses an eigendecomposition and a fixed ridge solve rather than fitting a full probabilistic observation model. Both the mean and basis are exchanged together. Calling this a basis transfer is therefore shorthand for transferring the complete fitted linear representation.

For a test image, sixty visible cells constrain sixteen latent coefficients. A ridge penalty of 0.01 stabilizes that visible-cell fit; its role is regularization, consistent with the motivation of Hoerl and Kennard, 1970. Predictions for the four concealed cells are then clipped to the interval from zero to one. The rank, penalty, clipping bounds, and concealed location are identical in both directions and are not selected using test errors. Fixed rules make the comparison interpretable, while also limiting it to one point in a much larger design space.

This setting shares the idea of recovering withheld content with learned inpainting, but not its representational capacity. Context-based inpainting Pathak et al., 2016 and irregular-hole partial convolution Liu et al., 2018 define richer reconstruction problems. Large-mask Fourier-based restoration Suvorov et al., 2021 and diffusion-based inpainting Lugmayr et al., 2022 supply additional methodological alternatives. Their relevance is to identify what this study leaves outside its scope: learned spatial hierarchies, complex masks, and generative sampling. None supplies an unmeasured benchmark score for our two fitted linear policies.

Other restoration approaches emphasize different sources of structure. Denoising autoencoders learn from corruption-based reconstruction tasks Vincent et al., 2008, whereas the deep image prior uses network structure as a restoration prior Ulyanov et al., 2017. State-space vision and restoration studies Liu et al., 2024 and Guo et al., 2024 show why an architectural comparison would require its own implementation and training record. Here, holding the inference algebra fixed makes the empirical contrast about the source of the fitted representation. It cannot reveal whether a more expressive model would remove, retain, or reverse the observed asymmetry.

3. Public resources, transformations, and the research model

The official MNIST partitions contain 60,000 training images and 10,000 test images. Each twenty-eight-by-twenty-eight image is normalized by 255 and converted to an eight-by-eight array using exact area-overlap averaging. An output cell spans 3.5 source pixels along each axis. The transformation uses the fractional overlap of those intervals, with weights that sum to one. It does not select a nearest source pixel or threshold the grayscale image. All MNIST comparisons in this article use that same transformed representation.

Optdigits supplies 3,823 training records and 1,797 test records, each with sixty-four features and a class label. The source documentation states that thirty writers contributed to training and thirteen different writers to testing. Its features count foreground pixels in nonoverlapping four-by-four blocks of normalized thirty-two-by-thirty-two bitmaps. We divide the resulting zero-to-sixteen counts by sixteen. The original compressed bitmap files are not used. Feature arrays are reshaped to eight-by-eight cells, but that common support does not undo the differing preprocessing histories of the two resources.

Using one generator initialized with seed 20261107, the script independently selects 2,000 training records from each official training partition. Labels do not enter selection weights, covariance estimation, latent fitting, or error calculation. The two subsets have equal size, without deliberate class balancing. Each resource supplies its own sample mean and covariance matrix. The sixteen eigenvectors associated with the largest eigenvalues form its basis. Numerical assertions verify orthonormality and the residual of every test-set ridge normal equation to a tolerance below one ten-billionth.

Rows three and four and columns three and four, using zero-based indexing, form the central two-by-two hidden region. The remaining sixty cells are visible. Every test image is reconstructed twice: once from the locally fitted representation and once from the other resource's representation. Squared error is averaged over only the four hidden cells. The mask is an artificial intervention on complete released examples, not naturally occurring archival damage. Missing-data theory helps distinguish a declared concealment mechanism from an observed pattern whose causes must be inferred Little and Rubin, 2002.

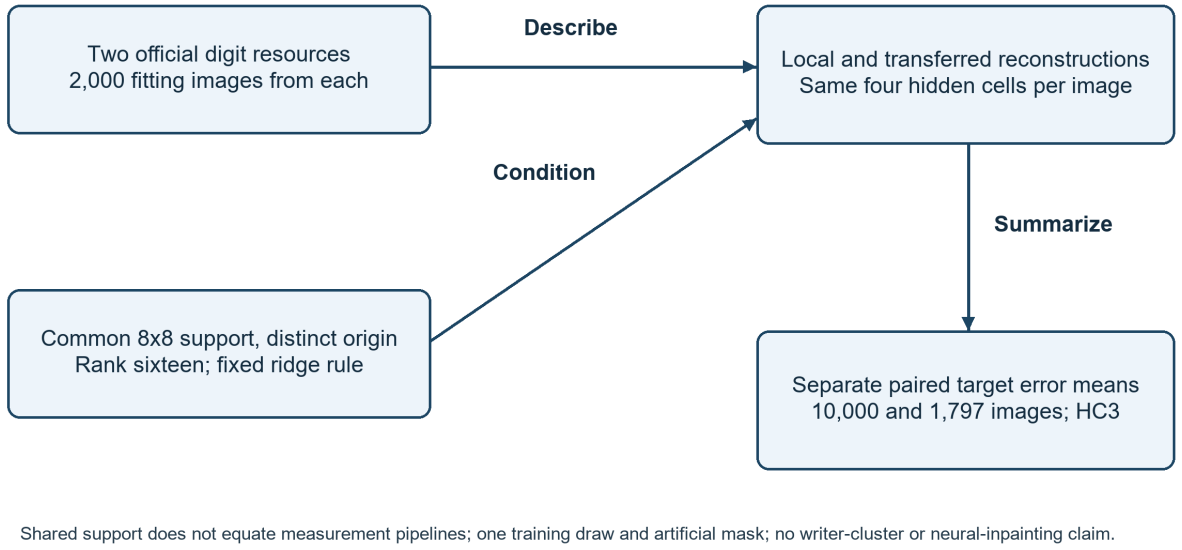


Figure 1. Equal fitting counts, source-specific representations, and paired target losses.

Figure 1 links the fitted source to the paired error comparison within each target resource. The observed-cell information, mask, rank, penalty, and clipping rule stay fixed within that comparison. Source preprocessing, writer composition, and the particular training draw change together. Measurement-error reasoning Carroll et al., 2006 cautions against treating numerical agreement of scales as proof that those underlying measurements are equivalent. The figure therefore contains no path assigning a causal effect to writer identity, binarization, or any single acquisition step.

4. Estimands and uncertainty

For each target image, the analysis subtracts local masked-cell mean squared error from transferred masked-cell mean squared error. The mean of that paired difference is the primary estimand. Separate intercept-only ordinary least-squares regressions estimate the MNIST and Optdigits contrasts. Pairing holds each target image and its concealed values constant across policies. It avoids comparing the absolute error levels of two different target resources as though those levels alone ranked intrinsic task difficulty or universal model quality.

The reported standard errors use HC3 heteroskedasticity adjustment, within the covariance-estimation tradition of White, 1980. Intervals use the normal approximation implemented for these fitted models.

There are 10,000 paired rows for MNIST and 1,797 for Optdigits. An intercept-only regression has no explanatory covariates, so its zero R-squared is a model-description field, not evidence that the paired contrast is absent. No class, image-complexity, or source-intensity control is added after inspecting the results.

These intervals condition on the two fitted representations and one mask. They do not incorporate uncertainty from drawing different training subsets. Nor do they supply writer-cluster inference. The fixed feature and IDX files used here do not identify the individual writer of each record. Cluster-robust methods require a defensible dependence grouping Cameron and Miller, 2015; inventing such a grouping from row order would not solve the problem. Structured validation principles Roberts et al., 2017 instead motivate reporting which dependence information is unavailable and limiting the claimed sampling population accordingly.

Test data are used to evaluate the prespecified computation, not to tune its rank or penalty. That separation addresses the model-selection issue discussed by Varma and Simon, 2006, without claiming that one frozen split provides a complete generalization audit. Distribution-shift research Ovadia et al., 2019 also distinguishes evaluating a transported model from assuming its uncertainty estimates remain trustworthy. We produce deterministic reconstructions and conditional error intervals, not calibrated probabilities for the missing cells. Probability-calibration methodology Guo et al., 2017 cannot turn the clipped predictions into validated confidence statements about restored content.

5. The local representations perform better in both directions

On MNIST test images, the local representation yields mean concealed-cell error 0.1047463. Substituting the Optdigits-fitted representation raises it to 0.1447552. The paired increase is 0.0400089, with standard error 0.0011604 and a 95% interval from 0.0377345 to 0.0422832. These values describe the fixed transformed MNIST target and the two realized fits. They show a clear local advantage within this comparison, without implying that every individual image benefits from the local policy.

The reverse direction has a larger penalty in the chosen units. On Optdigits test records, local error is 0.0974468 and error from the MNIST-fitted representation is 0.2561273. The difference is 0.1586805, with standard error 0.0043261 and interval from 0.1502015 to 0.1671595. Both estimates are positive. The larger Optdigits penalty does not establish that Optdigits is universally harder, because target distributions and representation formation differ. It shows that portability is directional for these two frozen reconstruction policies.

Table 1. Conditional paired transfer penalties for concealed-cell mean squared error

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.0400 (0.0012)	[0.0377, 0.0423]	10000
M1: fit	R-squared = 0.0000	-	-	10000
M2	Mean outcome/contrast	0.1587 (0.0043)	[0.1502, 0.1672]	1797
M2: fit	R-squared = 0.0000	-	-	1797

Model labels: M1 = MNIST transfer minus locally fitted MSE; M2 = Optdigits transfer minus locally fitted MSE.

Table 1 reports two intercept-only OLS models with HC3 standard errors, no covariate controls, no fixed effects, and normal-reference intervals. M1 evaluates the MNIST target; M2 evaluates the Optdigits target. N counts test images, and the outcome is transferred minus locally fitted mean squared error over four concealed normalized cells. Models

condition on a single training draw and do not include writer-level or training-resampling uncertainty. R-squared is zero by construction for these mean-only specifications.

Squared error is useful because the hidden cells have known numerical targets. It is not a legibility score, a classification success rate, or a probability score for a generated object. The broader scoring-rule framework Gneiting and Raftery, 2007 encourages aligning the loss with the object being estimated. Our loss evaluates conditional numerical reconstruction under a chosen mask. An archaeological interpretation, character transcription, or downstream catalog decision would require additional outcomes and evaluators. Small image-space error cannot substitute for those missing judgments.

6. What the asymmetry permits us to learn

The most direct interpretation is that a common grid and equal fitting counts did not make the fitted linear representations interchangeable. Several explanations remain compatible with the result. Binarized block occupancy and averaged grayscale intensity can distribute variance differently; normalization can change where strokes fall relative to the central mask; writer and class composition can change covariance directions. The experiment changes these factors jointly. Without a crossed acquisition and preprocessing study, allocating the penalty among them would be speculative.

The digit experiment should also remain distinct from transfer into a historical script collection. The oracle-character dataset described by Wang and Deng, 2024 and the Omniglot challenge discussed by Lake et al., 2019 illustrate why character resources differ in intended tasks and organization. Their existence does not make those images interchangeable with the digits analyzed here. A specialized restoration project would need source-specific masks, meaningful evaluation targets, and a partition that respects its own collection structure before claiming that this result predicts its performance.

A useful follow-up would repeat the complete fit over multiple independent training draws and report how much the transfer penalty varies. Another would apply alternative measurement transformations to the same underlying images, separating preprocessing from population differences. Influence diagnostics such as those introduced by Cook, 1977 could investigate whether a small number of training images disproportionately shape a fitted representation; no such diagnostic is reported as completed here. These proposals target concrete uncertainties left by the current design rather than adding unsupported mechanism claims to the observed mean differences.

The very small model-based p-values do not resolve these design limits. The interpretation recommended by Wasserstein and Lazar, 2016 places effect sizes, context, and the analysis process ahead of a threshold declaration. Here the scientifically useful evidence is the positive penalty in each target, its magnitude in the declared scale, and the fact that its uncertainty is conditional. Increasing the number of test rows alone would not identify the omitted writer dependence or distinguish binarization from other source characteristics.

7. Reproducibility and conclusion

The analysis retains the official source partition sizes, the Optdigits ZIP checksum, the source metadata, the transformation weights, the fitting seed, per-image errors, and structured regression outputs. Optdigits is documented under CC BY 4.0, with creator attribution retained. MNIST source attribution and repository conditions are retained separately. Dataset documentation should make such differences visible, following the motivation of Gebru et al., 2021. Machine-readable provenance and stable identifiers also serve the findability and reuse goals articulated by Wilkinson et al., 2016.

At a matched count of 2,000 fitting images, local linear reconstruction was better in both target resources under the central four-cell mask. That result answers the stated question for a deliberately limited computation. It gives no neural-inpainting ranking and no guarantee about restoration of meaningful inscriptions. The practical lesson is to test the numerical portability claim at the level where it is made: shared dimensions and sample counts can support a controlled comparison, while the acquisition process still determines what the fitted coordinates mean.

References

1. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
2. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.
3. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
4. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
5. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
6. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.
7. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
8. Alpaydin, E.; Kaynak, C. (1998). Optical Recognition of Handwritten Digits [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.
9. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.
10. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
11. Vincent, Pascal; Larochelle, Hugo; Bengio, Yoshua; Manzagol, Pierre-Antoine. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390294.
12. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.
13. Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. https://doi.org/10.1080/14786440109462720.
14. Lake, Brenden M.; Salakhutdinov, Ruslan; Tenenbaum, Joshua B.. (2019). The Omniglot challenge: a 3-year progress report. https://arxiv.org/abs/1902.03477.
15. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.
16. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
17. Liu, Yue; Tian, Yunjie; Zhao, Yuzhong; Yu, Hongtian; Xie, Lingxi; Wang, Yaowei; Ye, Qixiang; Jiao, Jianbin; Liu, Yunfan. (2024). VMamba: Visual State Space Model. arXiv:2401.10166. https://arxiv.org/abs/2401.10166.
18. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
19. Wang, Mei; Deng, Weihong. (2024). A dataset of oracle characters for benchmarking machine learning algorithms. Scientific Data, 11, 87. https://doi.org/10.1038/s41597-024-02933-w.
20. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.

21. Guo, Hang; Li, Jinmin; Dai, Tao; Ouyang, Zhihao; Ren, Xudong; Xia, Shu-Tao. (2024). MambaIR: A Simple Baseline for Image Restoration with State-Space Model. arXiv:2402.15648. https://arxiv.org/abs/2402.15648.

22. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

23. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

24. Ulyanov, Dmitry; Vedaldi, Andrea; Lempitsky, Victor. (2017). Deep Image Prior. arXiv:1711.10925. https://arxiv.org/abs/1711.10925.

25. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

26. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.

27. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

28. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE. https://doi.org/10.1109/5.726791.

29. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

30. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.