Journal of Integrated Knowledge
Vol. 2 No. 7 (2026)

Perceived Course Difficulty and the Gap between Course and Instructor Feedback

Logan Wilson[1] · Austin Anderson[2] · Ryan King[3]

[1] Denmark · [2] Norway · [3] Italy

Published: July 08, 2026 | Article 865

Abstract

The highest perceived-difficulty category showed little separation in how respondents scored instructor items relative to course items. In 5,820 released teaching-evaluation records, its adjusted gap relative to the lowest category was 0.040 coded points, with a working interval of −0.019 to 0.098. A second outcome based on upper-category response fractions gave 0.45 percentage points, also with an interval crossing zero. Intermediate difficulty categories had larger estimated contrasts. Course omissions preserved the coded-score contrast's direction but changed the fraction contrast's sign. These results describe paired questionnaire summaries and leave both teaching effectiveness and a causal difficulty effect unmeasured.

Perceived Course Difficulty and the Gap between Course and Instructor Feedback

Abstract

The highest perceived-difficulty category showed little separation in how respondents scored instructor items relative to course items. In 5,820 released teaching-evaluation records, its adjusted gap relative to the lowest category was 0.040 coded points, with a working interval of -0.019 to 0.098. A second outcome based on upper-category response fractions gave 0.45 percentage points, also with an interval crossing zero. Intermediate difficulty categories had larger estimated contrasts. Course omissions preserved the coded-score contrast's direction but changed the fraction contrast's sign. These results describe paired questionnaire summaries and leave both teaching effectiveness and a causal difficulty effect unmeasured.

Two objects inside one feedback form

A student can evaluate the provision of a course and the conduct of its instructor on the same questionnaire. The answers share an occasion, response scale and respondent, yet address different objects. A course might receive unfavorable comments about resources while its instructor receives favorable comments about explanations. Pooling every answer into one total would hide that relative evaluation. Comparing separately collected course and instructor samples would introduce a different problem: their respondents could differ. The released form studied here allows both blocks to be compared within each numerical record.

Our research question concerns the association between that paired block difference and perceived course difficulty, conditional on the recorded course, attendance and repetition categories. The prespecified main contrast compares difficulty category five with category one. Difficulty enters categorically because its steps need not represent equal increments. We also retain the other three category contrasts, allowing a pattern that rises and falls instead of imposing a linear difficulty gradient. The study estimates conditional associations in an archive, without attributing a change in feedback to an intervention that made a course harder.

Research on student evaluation has long distinguished dimensions of feedback and their intended uses (Marsh, 1984). The review by Spooren et al. (2013) organizes questionnaire design, dimensionality and potential bias around validity. Learning outcomes form a further criterion in the meta-analysis by Uttl et al. (2017), while Boring (2017) studies gender bias using information unavailable here. We do not turn a difference between questionnaire blocks into a finding about learning or discrimination. Those questions require observations beyond the released item responses and category codes.

The designated feedback study describes scale validation and uncertainty-aware experience evidence supporting decisions (Yuan et al., 2026). Its primary abstract motivates keeping a measurement convention separate from the decision it might inform. We do not access its private survey or implement its framework. Our narrower contribution is an explicit paired outcome, a fixed conditional model and a record of which uncertainty statements the public numerical support permits. Neither the presence of anonymous responses nor an interval plot establishes privacy protection.

Source records and the wording-defined comparison

The UCI Turkiye Student Evaluation archive provides 5,820 numerical records, 28 evaluation items and five contextual fields (Fokoue and Gunduz, 2013). We downloaded the complete official ZIP, retained both supplied CSV files, checked archive CRCs and recorded local SHA256 hashes. The current repository grants CC BY 4.0 reuse with attribution. The generic CSV contains no missing selected values. All item responses are integers from one through five; instructor codes have three values, course codes thirteen, attendance five, repetition three and perceived difficulty five.

This reuses the same complete numerical collection examined in an earlier project report about item covariance and constant response patterns. A direct hash comparison confirmed identical generic CSV bytes. The present analysis does not supply another university or an independent replication. Its question concerns a within-record block contrast and its conditional difficulty profile, rather than the source of item coherence. No correlation, alpha coefficient or previously reported effect is carried over as a new computed result.

We assign Q1-Q12 to the course block and Q13-Q28 to the instructor block using the documented wording. The former concern provision, resources, experience and benefit; the latter address instructor preparation, explanation, conduct and assessment. These are descriptive partitions. Measurement theory requires specifying how observations represent the intended construct (Hand, 1996), and validity concerns the inference drawn from an assessment (Messick, 1995). A convenient partition does not establish two latent traits, their independence or their suitability for a personnel decision.

There is no persistent student identifier, assessment date or enrollment roster. Consequently, a source row cannot be authenticated as a distinct independent student. The file contains 1,843 repeated complete numerical patterns beyond their first occurrence. We retain them: discrete questionnaires can produce equal patterns for different people, and the archive does not identify redundant submissions. The operational unit is the released response record. Documentation of that unit follows the provenance concern articulated in datasheets for datasets (Gebru et al., 2021).

The course codes also require care. Twelve occur with one instructor code each, whereas code thirteen occurs with two: 99 records for instructor two and 742 for instructor three. An initial program assertion of complete nesting failed before either outcome was constructed. We saved that failed check and corrected the interpretation in a dated support amendment. The original 23-column model was retained. Course-code adjustment therefore does not fully remove instructor composition, and no assertion of fully nested instructors remains.

Defining the two outcomes before fitting

For the main outcome, each record contributes its mean across sixteen instructor items minus its mean across twelve course items. Positive values indicate relatively higher instructor-item responses on the original coding. Each block mean uses equal item weights, despite different block lengths. Each released record then receives equal weight in regression. Weighting should follow the quantity sought (Solon et al., 2015); the resulting estimand is neither an item-count-weighted total nor a population estimate corrected by known sampling probabilities.

The main gap lies between minus four and four coded points. Its numerical meaning depends on the released one-to-five coding. Stevens (1946) distinguishes permissible interpretations of measurement scales, and Liddell and Kruschke (2018) show why treating ordinal measurements as metric can distort substantive conclusions. We therefore interpret this result as a difference between specified coded

summaries. It is not an estimated distance between latent satisfaction states. Averaging many items does not independently justify equal psychological spacing.

A second, prespecified outcome replaces each item by an indicator that its response belongs to categories four or five. We subtract the course block's indicator fraction from the instructor block's fraction. The result ranges from minus one to one and uses the same records and predictors. Multiplication by one hundred expresses it in percentage points. This comparison depends on the chosen upper-category boundary, but not on the distances among the five original codes. It remains a separate operational outcome rather than a correction proving the first scale valid.

Ordinal regression offers a different route when the research target is a category distribution (Bürkner and Vuorre, 2019). We do not estimate such a model here or compare several thresholds after inspecting their outcomes. Our two summaries were declared before effects were calculated. Bollen and Lennox (1991) emphasize that measurement assumptions determine what a composite represents; Jacobs and Wallach (2021) likewise connect operational choices to subsequent interpretations. Neither composite should quietly acquire the meaning of overall teaching quality.

Paired subtraction can remove a common additive response level from the two means. It cannot eliminate item interpretation, differential scale use or the unequal content of the blocks. Response strategies discussed by Krosnick (1991) and common-method concerns reviewed by Podsakoff et al. (2003) are therefore relevant limits. We apply no rule that labels a uniform pattern careless, authentic or invalid. A row giving the same category to all items has a zero block gap by construction and remains part of the declared released-record target.

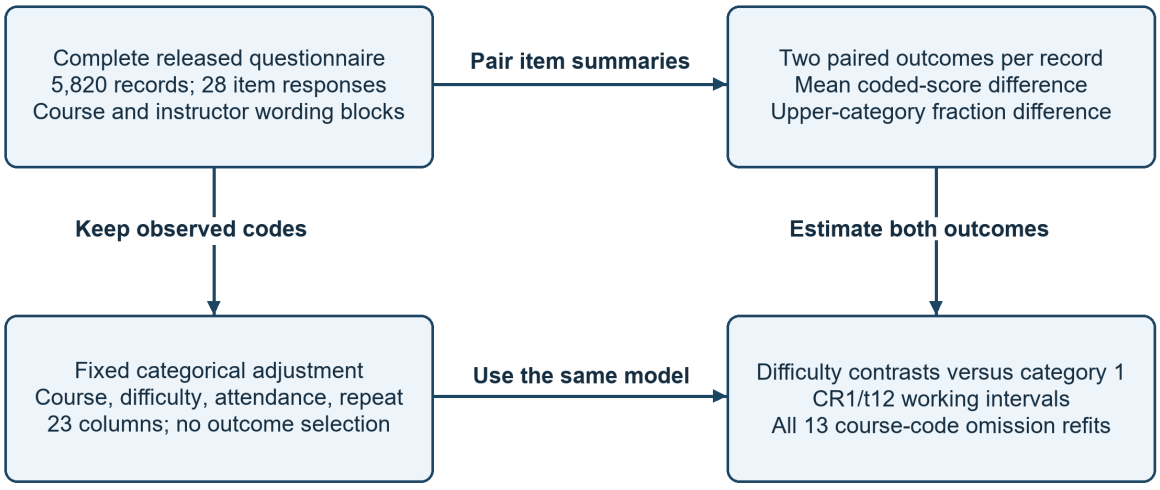

Figure 1. The analysis compares wording-defined feedback summaries. Difficulty, attendance, repetition and course identifiers are observed categories. The arrows specify a statistical calculation and provide no causal or latent-scale interpretation.

Conditional models and their uncertainty limits

Both outcomes use ordinary least squares with an intercept, twelve course-code indicators, four difficulty indicators, four attendance indicators and two repetition indicators. Reference categories are respectively course one, difficulty one, attendance zero and repetition one. There are 23 columns and 5,820 observations. All difficulty categories occur within every course code, so the extreme contrast is not based on a completely absent course-difficulty cell. Nevertheless, the number of records in those cells varies substantially.

The fixed-effects interpretation follows the conditional linear-model framework described by Wooldridge (2010). The coefficient on difficulty five is the fitted gap difference relative to difficulty one under the stated additive adjustment. It is not the effect of increasing an individual's difficulty response while holding every unmeasured determinant constant. Instructors, students and course conditions were not randomized. Reported difficulty can itself be associated with expectations and evaluations, and the release supplies no time ordering to resolve that relationship.

Working uncertainty uses a course-code clustered sandwich covariance, with the finite correction G divided by G minus one and N minus one divided by N minus K. Here G is thirteen and K is 23. Intervals use the Student t reference with twelve degrees of freedom. Cluster-robust inference depends on its clustering and replication assumptions (Cameron and Miller, 2015). Repeated observations do not become independent merely because their within-group covariance is handled in a calculation; this distinction is also central to longitudinal modeling (Liang and Zeger, 1986).

Thirteen codes provide limited support for an asymptotic cluster approximation. Small-sample fixed-effects methods address that difficulty (Pustejovsky and Tipton, 2018), but we have not substituted their correction while retaining a different estimator's label. The reported intervals are the explicit CR1/t12 working calculation. Students may appear across course codes without linkage, and instructors span several codes. Such cross-code dependence remains unaddressed. Robust standard errors also cannot repair a wrong substantive model or provide causal identification (Freedman, 2006).

We additionally remove each course code and refit both models, preserving all thirteen omissions. Their minimum and maximum main contrasts describe finite composition sensitivity. They are not another confidence interval, and a positive omission range cannot establish population certainty. The protocol and source-support amendment precede the first outcome calculation. Advance specification helps distinguish fixed questions from discoveries selected during analysis (Wagenmakers et al., 2012), although this archived-data study is not externally preregistered or outcome-blind.

The extreme contrast is small and the profile is nonmonotonic

Across all records, the mean coded gap was 0.170819 points and the mean upper-category gap was 6.335911 percentage points. Those released-record averages combine course codes and difficulty groups. They do not themselves estimate the conditional contrast. Table 1 reports every difficulty coefficient from the two fixed models, retaining intervals crossing zero as well as positive point estimates. Each interval is pointwise; the table is not a family of independently confirmed discoveries.

Table 1. Adjusted difficulty contrasts in the two paired block gaps.

Difficulty versus 1	Coded gap, points	Working 95% interval	Upper-category gap, pp	Working 95% interval
2 versus 1	0.085508	[0.045565, 0.125450]	4.475661	[2.153547, 6.797775]
3 versus 1	0.064276	[0.025368, 0.103183]	3.299345	[0.934992, 5.663698]
4 versus 1	0.061552	[-0.003793, 0.126897]	2.738189	[0.018536, 5.457841]
5 versus 1	0.039953	[-0.018517, 0.098424]	0.451502	[-1.708255, 2.611259]

Table 1 note: Each model uses all 5,820 records, 23 columns and thirteen course-code clusters. Working95% intervals use CR1 and t12. The reference is difficulty one, with course, attendance and repetition categories adjusted. Coded points and percentage points are different outcome units. No instructor-quality or causal interpretation is attached to these coefficients.

The main difficulty-five contrast was 0.039953 coded points, with working interval -0.018517 to 0.098424. Under the upper-category definition it was 0.451502 percentage points, with interval -1.708255 to 2.611259. The available calculation therefore does not support a precise positive extreme-category separation. It also does not establish equivalence: no equivalence margin was declared and both intervals contain values on either side of zero.

Difficulty two had the largest point contrast in both outcomes: 0.085508 coded points and 4.475661 percentage points. Categories three and four lay between that estimate and the extreme-category estimate. This observed profile is compatible with more than one underlying response process. It provides a concrete reason to retain categorical difficulty rather than describe the results as a steadily increasing gap. The prespecified extreme comparison remains primary; the more favorable intermediate estimate does not replace it.

The difference between the two outcome definitions is also informative. A small shift among lower response categories can affect coded means without changing an upper-category indicator. A shift from three to four changes the latter, while one from four to five does not. We have not observed an underlying continuous attitude that chooses between these rules. The estimates should be read alongside their response definitions, rather than merged into a generic claim that instructor satisfaction increases with difficulty.

Course omissions and what the result can support

Table 2. Extreme-category contrast under all course-code omissions.

Outcome	Complete model	Omission minimum	Omission maximum	Negative omissions
Coded gap, points	0.039953	0.013512	0.051564	0 of 13
Upper-category gap, pp	0.451502	-0.447047	1.094337	1 of 13

Table 2 note: Each of thirteen source course codes is omitted once and the remaining records refitted. Ranges describe the fixed collection's composition sensitivity. They are not sampling intervals; the same respondents may be represented across unknown source occasions.

All thirteen coded-score omission estimates were positive, ranging from 0.013512 to 0.051564 points. The upper-category range was -0.447047 to 1.094337 percentage points and crossed zero. The ordinal-summary direction is therefore sensitive to which recorded course code is retained. This does not contradict the coded-score range: the outcomes assign different consequences to movements within

response categories. Neither omission exercise changes the unrecovered identities or supplies additional independent courses.

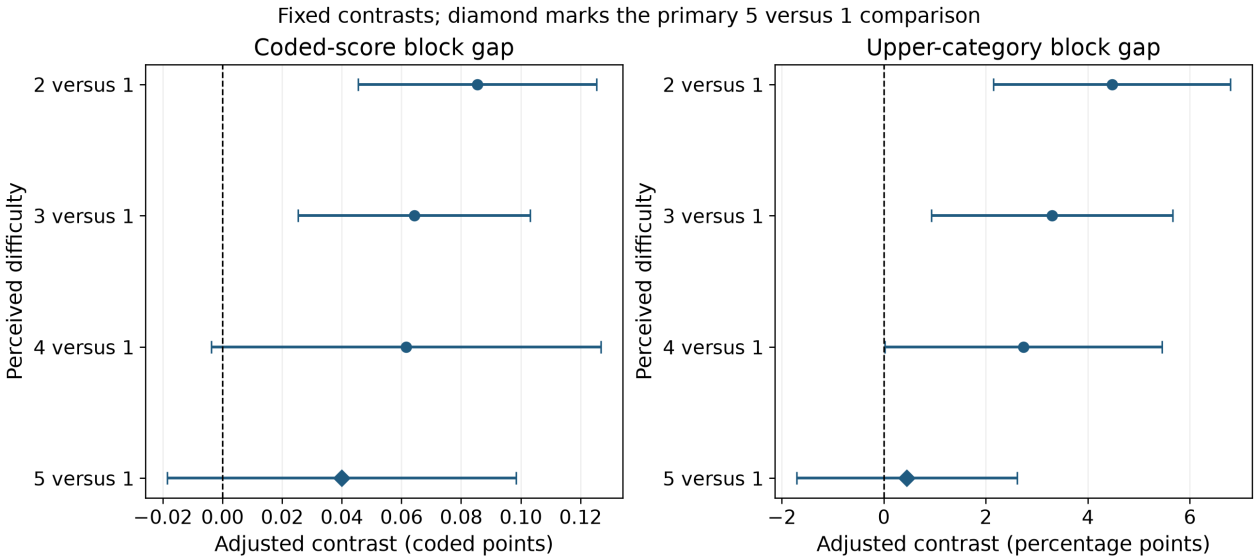

Figure 2. Dots are the four fixed category contrasts against difficulty one; bars show the stated CR1/t12 working intervals. The vertical zero line is shared conceptually, while each panel retains its own outcome unit. Intermediate and extreme estimates are all displayed, without a selected-sign axis or significance stars.

Both complete designs had full column rank. No fitted gap left its theoretical bounds under either outcome. That numerical check does not validate the linear conditional expectation or remove residual confounding. The complete paired outcome reconstruction, two full regressions, covariance matrices and all 26 omission coefficients were independently verified from the original CSV. An alternate QR solve and a separate software implementation agreed with the primary results within numerical tolerance. These checks establish calculation consistency, without resolving the archive's sampling limitations.

The practical contribution is a more specific description of the released feedback. Instructor-item responses exceed course-item responses on average, but the highest difficulty category does not exhibit a sharply estimated additional relative advantage. The intermediate profile and the upper-category omission reversal discourage a single linear story. A decision about instructor performance would still require external criteria, identifiable observation units and a justified measurement model. None is supplied by the paired difference alone.

Computational provenance

Primary arrays use NumPy (Harris et al., 2020), tabulation uses pandas (McKinney, 2010), and Student quantiles and the independent linear-algebra path use SciPy (Virtanen et al., 2020). Statsmodels supplies a separate coefficient and cluster-covariance check (Seabold and Perktold, 2010). Matplotlib creates the numerical figure from saved results (Hunter, 2007). The retained package includes source hashes, the original failed nesting assertion, its support amendment, fixed protocol, derived records, full coefficients, omissions and independent-check tolerances.

This audit trail follows the reproducible-computation aim discussed by Peng (2011). Procedural research assistance used the Scientific Agent Skills workflow (Kassis et al., 2026); it did not generate observations or supply an independent human judgment of the conclusions. Final source-claim and author review remain pending. The released files support checking the stated numerical comparison, with the sampling, measurement and instructor-composition limits preserved.

References

1. Colin Cameron, A.; Miller, Douglas L.. (2015). A Practitioner’s Guide to Cluster-Robust Inference. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.317
2. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677
3. Pustejovsky, James E.; Tipton, Elizabeth. (2018). Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2016.1247004
4. Bürkner, Paul-Christian; Vuorre, Matti. (2019). Ordinal Regression Models in Psychology: A Tutorial. Advances in Methods and Practices in Psychological Science. https://doi.org/10.1177/2515245918823199
5. Marsh, Herbert W.. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility.. Journal of Educational Psychology. https://doi.org/10.1037/0022-0663.76.5.707
6. Wagenmakers, Eric-Jan; Wetzels, Ruud; Borsboom, Denny; van der Maas, Han L. J.; Kievit, Rogier A.. (2012). An Agenda for Purely Confirmatory Research. Perspectives on Psychological Science. https://doi.org/10.1177/1745691612463078
7. Uttl, Bob; White, Carmela A.; Gonzalez, Daniela Wong. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation. https://doi.org/10.1016/j.stueduc.2016.08.007
8. Seabold, Skipper; Perktold, Josef. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-011
9. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumé, Hal; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723.
10. Liddell, Torrin M.; Kruschke, John K.. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. Journal of Experimental Social Psychology. https://doi.org/10.1016/j.jesp.2018.08.009
11. Freedman, David A. (2006). On The So-Called “Huber Sandwich Estimator” and “Robust Standard Errors”. The American Statistician. https://doi.org/10.1198/000313006x152207
12. LIANG, KUNG-YEE; ZEGER, SCOTT L.. (1986). Longitudinal data analysis using generalized linear models. Biometrika. https://doi.org/10.1093/biomet/73.1.13
13. Krosnick, Jon A.. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology. https://doi.org/10.1002/acp.2350050305
14. Fokoue, Ernest; Gunduz, Necla. (2013). Turkiye Student Evaluation. UCI Machine Learning Repository dataset. https://doi.org/10.24432/C5S02S
15. Kassis, Timothy; Agarwal, Vinayak; He, Yuhuan; Patel, Darshil; Brueckner, Aubrey M.. (2026). Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv preprint. https://arxiv.org/abs/2609.00065
16. Spooren, Pieter; Brockx, Bert; Mortelmans, Dimitri. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. Review of Educational Research. https://doi.org/10.3102/0034654313496870
17. Yuan, Yuewei; Huang, Jialei; Yin, Jiayang; Xu, Tianyi. (2026). Visual Verification of user Experience Differences under Privacy Constraints in Anonymous Feedback Data. SSRN working paper 7393298. https://doi.org/10.2139/ssrn.7393298.
18. Hunter, John D.. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55.
19. Hand, D. J.. (1996). Statistics and the Theory of Measurement. Journal of the Royal Statistical Society. Series A (Statistics in Society). https://doi.org/10.2307/2983326
20. Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning.. American Psychologist. https://doi.org/10.1037/0003-066x.50.9.741
21. Jacobs, Abigail Z.; Wallach, Hanna. (2021). Measurement and Fairness. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445901
22. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.

23. Virtanen, Pauli; Gommers, Ralf; Oliphant, Travis E.; Haberland, Matt; Reddy, Tyler; Cournapeau, David; Burovski, Evgeni; Peterson, Pearu; Weckesser, Warren; Bright, Jonathan; van der Walt, Stéfan J.; Brett, Matthew; Wilson, Joshua; Millman, K. Jarrod; Mayorov, Nikolay; Nelson, Andrew R. J.; Jones, Eric; Kern, Robert; Larson, Eric; Carey, C J; Polat, İlhan; Feng, Yu; Moore, Eric W.; VanderPlas, Jake; Laxalde, Denis; Perktold, Josef; Cimrman, Robert; Henriksen, Ian; Quintero, E. A.; Harris, Charles R.; Archibald, Anne M.; Ribeiro, Antônio H.; Pedregosa, Fabian; van Mulbregt, Paul; SciPy 1.0 Contributors; Vijaykumar, Aditya; Bardelli, Alessandro Pietro; Rothberg, Alex; Hilboll, Andreas; Kloeckner, Andreas; Scopatz, Anthony; Lee, Antony; Rokem, Ariel; Woods, C. Nathan; Fulton, Chad; Masson, Charles; Häggström, Christian; Fitzgerald, Clark; Nicholson, David A.; Hagen, David R.; Pasechnik, Dmitrii V.; Olivetti, Emanuele; Martin, Eric; Wieser, Eric; Silva, Fabrice; Lenders, Felix; Wilhelm, Florian; Young, G.; Price, Gavin A.; Ingold, Gert-Ludwig; Allen, Gregory E.; Lee, Gregory R.; Audren, Hervé; Probst, Irvin; Dietrich, Jörg P.; Silterra, Jacob; Webber, James T; Slavič, Janko; Nothman, Joel; Buchner, Johannes; Kulick, Johannes; Schönberger, Johannes L.; de Miranda Cardoso, José Vinícius; Reimer, Joscha; Harrington, Joseph; Rodríguez, Juan Luis Cano; Nunez-Iglesias, Juan; Kuczynski, Justin; Tritz, Kevin; Thoma, Martin; Newville, Matthew; Kümmerer, Matthias; Bolingbroke, Maximilian; Tartre, Michael; Pak, Mikhail; Smith, Nathaniel J.; Nowaczyk, Nikolai; Shebanov, Nikolay; Pavlyk, Oleksandr; Brodtkorb, Per A.; Lee, Perry; McGibbon, Robert T.; Feldbauer, Roman; Lewis, Sam; Tygier, Sam; Sievert, Scott; Vigna, Sebastiano; Peterson, Stefan; More, Surhud; Pudlik, Tadeusz; Oshima, Takuya; Pingel, Thomas J.; Robitaille, Thomas P.; Spura, Thomas; Jones, Thouis R.; Cera, Tim; Leslie, Tim; Zito, Tiziano; Krauss, Tom; Upadhyay, Utkarsh; Halchenko, Yaroslav O.; Vázquez-Baeza, Yoshiki. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. https://doi.org/10.1038/s41592-019-0686-2.

24. McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-00a.

25. Wooldridge, Jeffrey M. (2010). Econometric Analysis of Cross Section and Panel Data. Second edition. MIT Press. https://mitpress.mit.edu/9780262232586/econometric-analysis-of-cross-section-and-panel-data/

26. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847.

27. Boring, Anne. (2017). Gender biases in student evaluations of teaching. Journal of Public Economics. https://doi.org/10.1016/j.jpubeco.2016.11.006

28. Bollen, Kenneth; Lennox, Richard. (1991). Conventional wisdom on measurement: A structural equation perspective.. Psychological Bulletin. https://doi.org/10.1037/0033-2909.110.2.305

29. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.

30. Podsakoff, Philip M.; MacKenzie, Scott B.; Lee, Jeong-Yeon; Podsakoff, Nathan P.. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies.. Journal of Applied Psychology. https://doi.org/10.1037/0021-9010.88.5.879