# Managing lexical-semantic hybrid records of FAIR Metrics analyses with the NPDS Cyberinfrastructure

Adam Craig, Anousha Athreya, Carl Taswell

Brain Health Alliance, Ladera Ranch, CA, USA

BHAVI Symposium online 9 October 2023

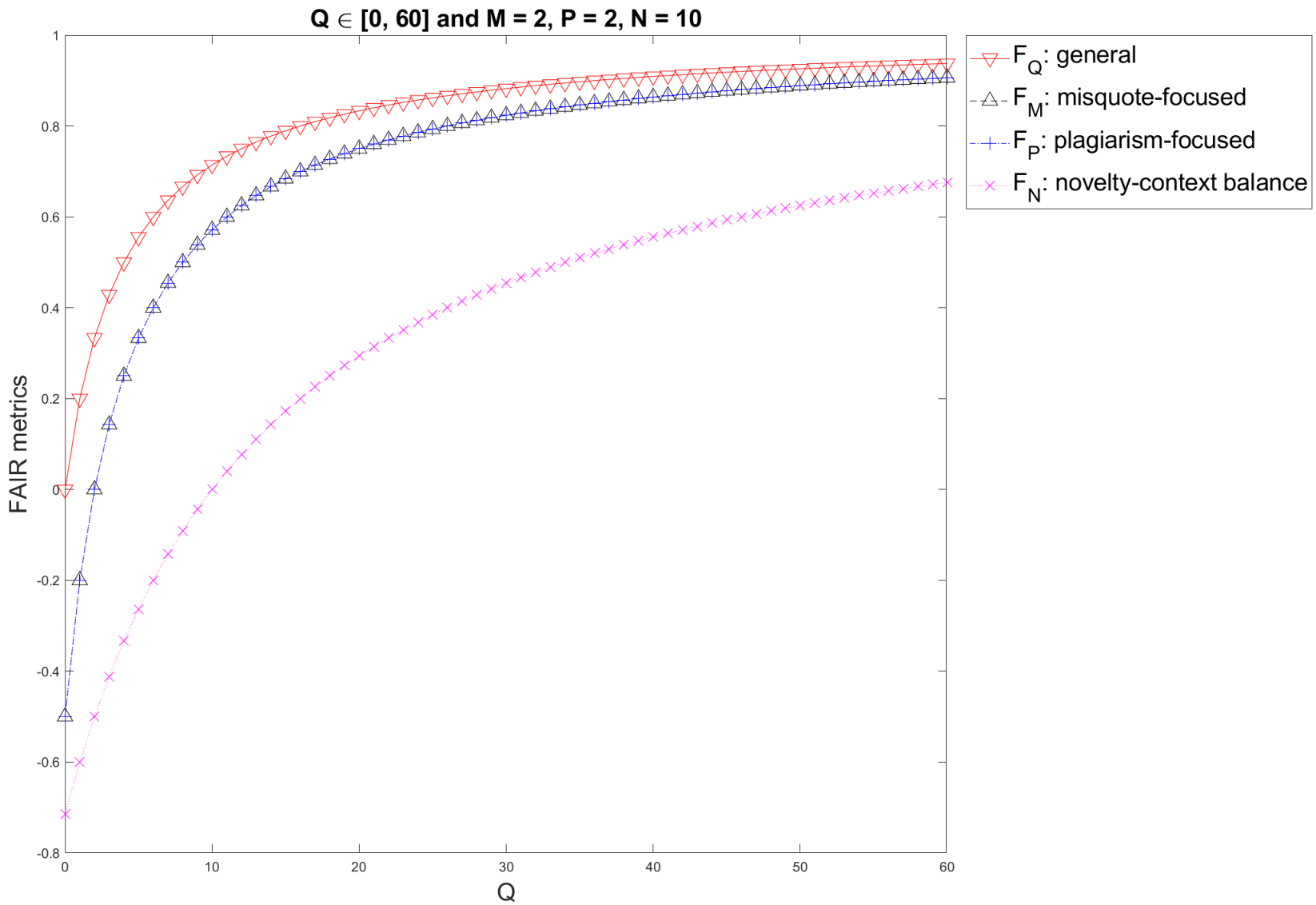

# The need for objectivity

- Authors may avoid citing the work of potential rivals.
- They may also misrepresent the content of prior work.
- Peer reviewers and editors may have their own biases or perverse incentives.
- Institutional ethics committees may care more about avoiding damage to the institute's reputation than about righting wrongs.
- See (Taswell et al., 2020, *ASIS&T 2020*) for a review of these issues.
- We need an alternative to subjective judgments: **Quantify it.**
- In (Craig & Taswell, 2018, *ASIS&T-SIGMET 2018*), we proposed FAIR Attribution to Indexed Reports (FAIR) Metrics of adherence to good citation practices.
- In the present work, a human evaluator demonstrates their use with 5 published articles from scholarly journals.

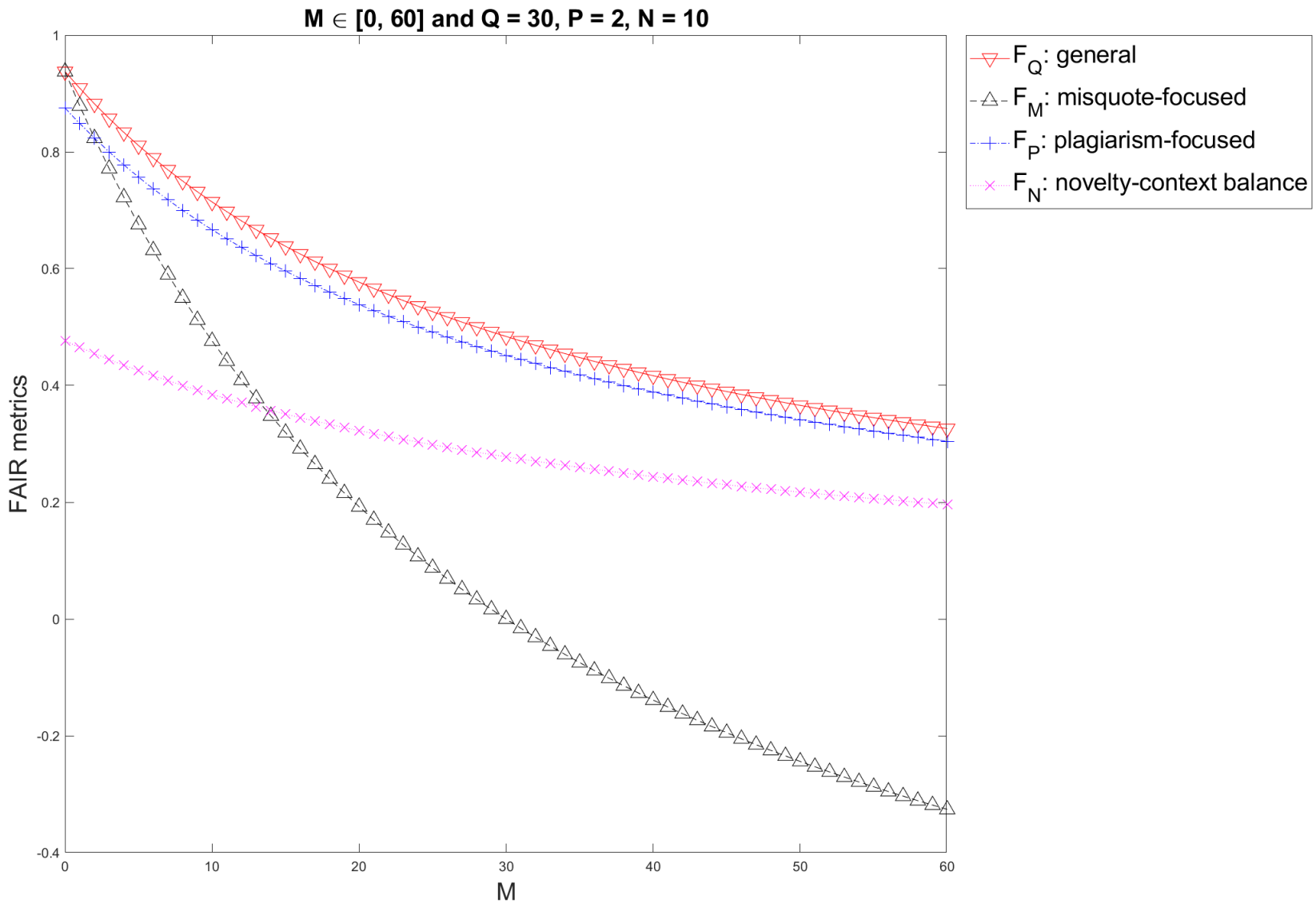
# The FAIR Metrics

- **In (Craig & Taswell, 2018, *BIBM 2018*), we introduced 4 counts of 4 categories of claims.**
- *Q*uoted: statements correctly attributed to prior work
- *M*isquoted: statements misrepresenting the content of prior work
- *P*lagiarized: statements presented as novel but found in prior work
- *N*ovel: statements presented as novel and not found in prior work
- **In (Craig et al., 2019, *ASIS&T 2019*), we introduced 4 ratio FAIR Metrics, each with a different emphasis.**
- $F_Q = \frac{Q}{Q+P+M}$: overall frequency of valid attributions to prior work
- $F_M = \frac{Q-M}{Q+P+M}$: emphasis on misrepresentation
- $F_P = \frac{Q-P}{Q+P+M}$: emphasis on plagiarism
- $F_N = \frac{Q-N}{Q+P+M+N}$: balance of new ideas vs context from prior work

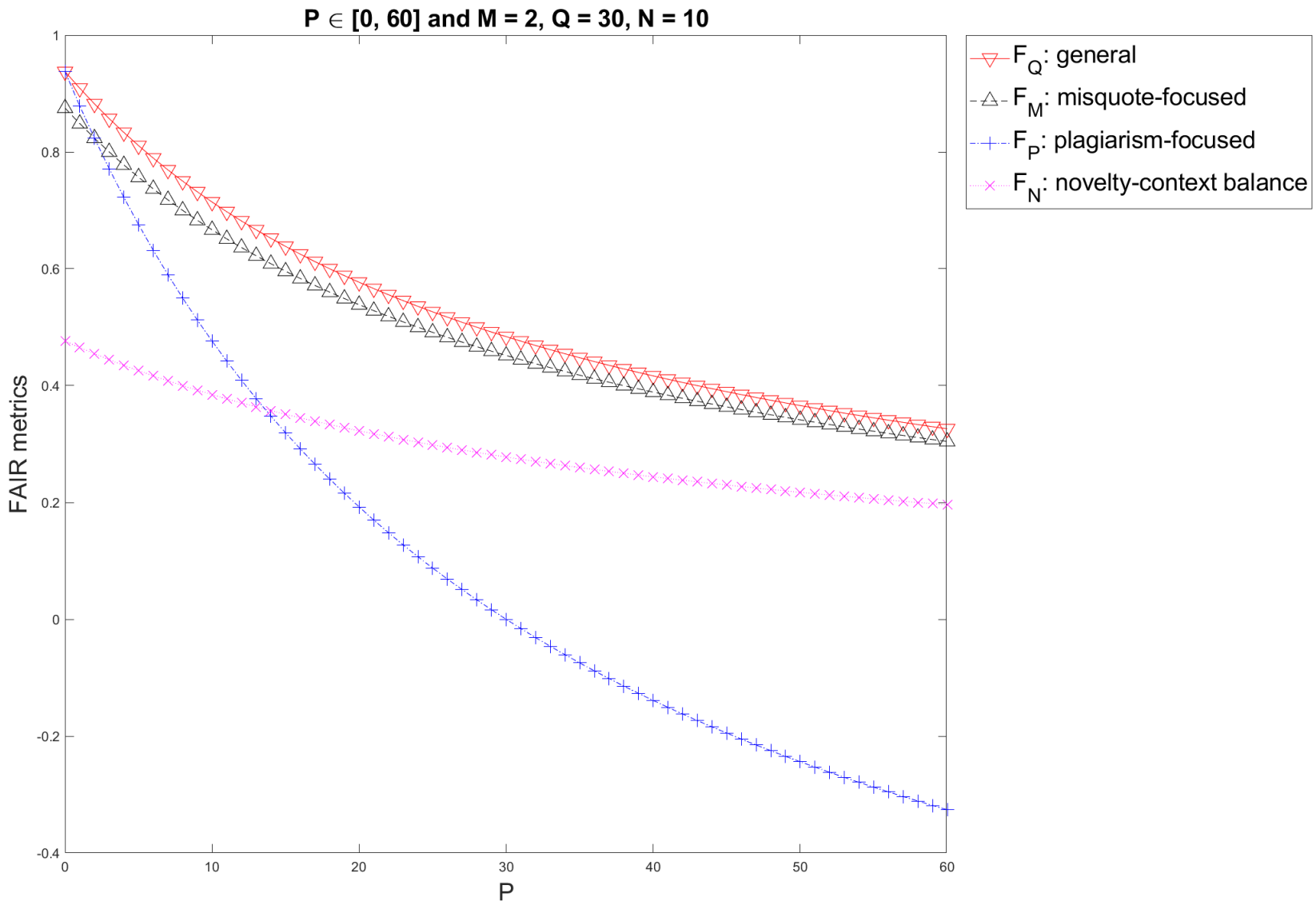
# Behaviors of the FAIR Metrics with increasing $Q$



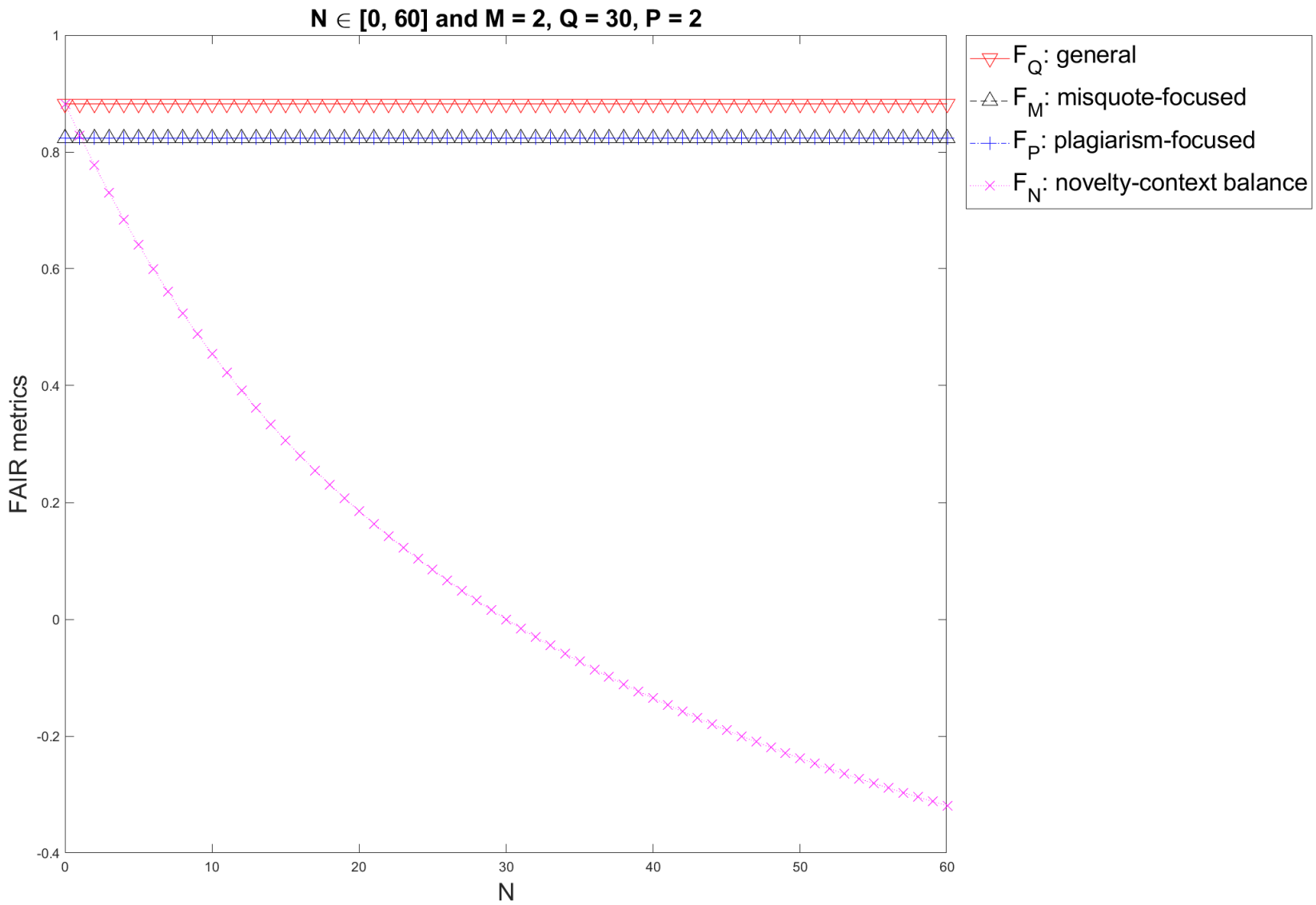
# Behaviors of the FAIR Metrics with increasing $M$



# Behaviors of the FAIR Metrics with increasing $P$



# Behaviors of the FAIR Metrics with increasing $N$



# Target evaluation by human reviewer

- Review a Target text, $T$, in comparison to a Comparison text, $C$.
- Identify the key claims of $T$.
- For each claim attributed to a prior work, search that prior work for an equivalent claim.
  - If found, count the claim in $T$ as *Q*uoted.
  - If not, count the claim in $T$ as *M*isquoted.
- For each claim presented as novel, search $C$ for an equivalent claim.
  - If found, count the claim in $T$ as *P*otentially Plagiarized.
  - If not, count the claim in $T$ as *N*ovel.

# Standard of equivalence

- For a detailed discussion of different interpretations of "equal or equivalent entities," see (Athreya, 2020, *TransAI 2020*).
- When comparing a statement in the Target work, $A$, to a statement in a prior work, $B$, assign the match a score of -1 to 4.
- 4: $A$ and $B$ are an exact lexical match.
- 3: $A$ is a close paraphrasing of $B$.
- 2: $A$ is a reasonable summary of $B$.
- 1: $A$ has some information from $B$ but also adds to it.
- 0: $A$ and $B$ are clearly different in meaning.
- -1: $A$ contradicts $B$.
- Count any score of 2 or higher as equivalent.

# Claims vs Statements

- For the present study, we focus on the key Claims of an article, not every Statement in it.
- For our purposes, a Statement is any assertion of fact.
- Claims are Statements that are significant to the main argument that the article is making.
- Valid claims can be Novel observations and insights or Quoted from prior work.

# FAIR Metrics results

| **Target text** | **Ret--racted?** | **Comparison text** | $M$ | $N$ | $P$ | $Q$ | $F_M$ | $F_N$ | $F_P$ | $F_Q$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Taswell 2007 | no | Mons 2005 | 0 | 20 | 0 | 22 | 1.00 | 0.05 | 1.00 | 1.00 |
| Uddin 2022 | yes | Foster et al. 2019 | 0 | 18 | 18 | 87 | 0.83 | 0.56 | 0.66 | 0.83 |
| Gnat et al. 2022 | yes | de Hoog et al. 2017 | 0 | 3 | 10 | 30 | 0.75 | 0.63 | 0.50 | 0.75 |
| Ullah et al. 2018 | yes | Sansaniwal & Kumar 2015 | 31 | 3 | 7 | 2 | -0.73 | -0.02 | -0.13 | 0.05 |
| Wilkinson et al. 2016 | no | Taswell 2007 | 6 | 5 | 24 | 28 | 0.38 | 0.37 | 0.07 | 0.48 |

- Target: the text for which we are calculating FAIR Metrics.
- Retracted?: Was Target retracted for plagiarism of Comparison?
- Comparison: We are checking the Target for plagiarism of this text.
- $M, N, P, Q$ Counts: *M*isquoted, *N*ovel, *P*lagiarized, *Q*uoted.
- $F_M = \frac{Q-M}{Q+P+M}$; $F_N = \frac{Q-N}{Q+P+M+N}$; $F_P = \frac{Q-P}{Q+P+M}$; $F_Q = \frac{Q}{Q+P+M}$

# Case 1: Taswell 2008 vs Mons 2005

- *C* Mons, B. (2005). Which gene did you mean?. *BMC bioinformatics*, 6(1), 1-4.
- *T*: Taswell, C. (2008). DOORS to the semantic web and grid with a PORTAL for biomedical computing. *IEEE Transactions on Information Technology in Biomedicine*, 12(2), 191-204.
- First publication describing the PORTAL and DOORS service types
- Received 2006-10-31, revised 2007-06-11, and published 2008-03-05
- Works have very different emphases:
- Mons describes at length why semantic markup is important.
- Taswell mostly cites other authors' commentaries on this.
- Mons has a clear focus on high-throughput experiments in genetics/molecular biology.
- Taswell speaks broadly of biomedical computing and of cross-domain utility.
- Mons discusses who should be creating semantic markup.
- Taswell discusses how best to manage and disseminate it.

# Case 2: Uddin et. al 2020 vs Foster et al. 2019

- *C*: Foster, Evangeline M., Adrià Dangla-Valls, Simon Lovestone, Elena M. Ribe, & Noel J. Buckley. Clusterin in Alzheimer's disease: mechanisms, genetics, and lessons from other pathologies. *Frontiers in neuroscience* 13 (2019): 164.
- *T*: Uddin, M., Kabir, M., Begum, M., Islam, M., Behl, T., & Ashraf, G. M. (2021). Exploring the Role of CLU in the Pathogenesis of Alzheimer's Disease. *Neurotoxicity Research*, 39(6), 2108-2119.
- Review of work investigating how clusterin (aka. apolipoprotein J) has both neuroprotective and neurotoxic roles in Alzheimer's Disease
- Received 2020-06-04, revised 2020-08-05, accepted 2020-08-10, published 2020-08-21, corrected 2020-08-29, **retracted 2022-03-19**
- Udin et al. do add some new content.
- FAIR Metrics designed for primary research articles.
- But review $\neq$ list of quotes.
- Can measure plagiarism of synthesis, commentary.

# Case 3: Gnat et al. 2022 vs de Hoog et al. 2017

- *C*: de Hoog, G. S., Dukik, K., Monod, M., Packeu, A., Stubbe, D., Hendrickx, M., ... & Gräser, Y. (2017). Toward a novel multilocus phylogenetic taxonomy for the dermatophytes. *Mycopathologia*, 182(1), 5–31.
- *T*: Gnat, S., Nowakiewicz, A., & Zięba, P. (2019). Taxonomy of dermatophytes–the classification systems may change but the identification problems remain the same. *Postępy Mikrobiologii-Advancements of Microbiology*, 58(1), 49-58.
- Received 2018-08-01, accepted 2018-11-01, published 2019-06-10, **retracted 2022-10-30**
- Unusually, they cite de Hoog et al. but also have numerous unattributed statements paraphrased from it.

# Case 4: Ullah et al. 2018 vs Sansaniwal & Kumar 2015

- *C*: Sansaniwal, S. K., & Kumar, M. (2015). Analysis of ginger drying inside a natural convection indirect solar dryer: An experimental study. *Journal of Mechanical Engineering and Sciences*, 9, 1671-1685.
- *T*: Ullah, F., Kang, M., Khattak, M. K., & Wahab, S. (2018). Retracted: Experimentally investigated the asparagus (*Asparagus officinalis L.*) drying with flat-plate collector under the natural convection indirect solar dryer. *Food Science & Nutrition*, 6(6), 1357-1357.
- Received 2017-11-23, revised 2018-01-08, accepted 2018-01-10, published 2018-02-21, **retracted 2018-09-19**
- Ullah et al. copied nearly the entire paper from Saniswal & Kumar.
- They then replaced "ginger" with "asparagus" as the vegetable being dried in the solar dryer.
- Changing a content word decreased the number of equivalent claims.
- Some Plagiarized became Novel.
- Far more Quoted became Misquoted.

# Case 5: Wilkinson et. al 2016 vs Taswell 2007

- *C*: Taswell, C. (2007). DOORS to the semantic web and grid with a PORTAL for biomedical computing. *IEEE Transactions on Information Technology in Biomedicine*, 12(2), 191-204.
- *T*: Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.
- Received 2015-12-10, accepted 2016-02-12, and published 2016-03-15, not (yet) retracted
- (Craig et al., 2019, *ECAI 2019*) details how all "FAIR Principles" are equivalent to pre-existing design principles from the PORTAL-DOORS Project.
- # of statements about motivation and community-building $\gg$ # of statements about the principles
- Novel statements are from the former set.

# The PDP-DREAM Ontology

- Claims can have unique identifiers: <pdpd:UniqueIdentifierPrinciple>
- We can also represent them as triples: <pdpd:LabelServerRecord> <pdpd:hasMetadata> <pdpd:UniqueURIOrIRI> .
- Original text as a property: <pdpd:UniqueIdentifierPrinciple> <pdpd:hasText> “Thus, resource label servers (as the analogs in DOORS of the domain name servers in DNS) should maintain database records with the following required metadata for each resource: 1) the resource label with a globally unique URI (or IRI) enabling nonsemantic string queries of labels...” .
- Match as a property: <pdpd:UniqueIdentifierPrinciple> <pdpd:hasEquivalent> “To be Findable: ... F1. (meta)data are assigned a globally unique and persistent identifier.” .
- FAIR Metric value as a property: <fidentinus:Wilkinson2016FAIRGPSD> <pdpd:hasFAIRF3Value> “0.47” .

# Resource Records and Diristries

- Uddin et. al 2020, Gnat et al. 2022, Ullah et al. 2018, and Wilkinson et al. 2016 placed in `www.portaldoors.org` → **Fidentinus** diristry for known or suspected **plagiarism cases**.
- Non-plagiarizing papers placed according in diristry with best match for problem domain.
- Taswell 2008, Mons 2005 placed in `www.portaldoors.org` → **DaVinci** diristry for **semantic web** resources
- Foster et al. 2019 placed in `brainwatch.net` → **SOLOMON** diristry for hypotheses about diseases causing **neurodegeneration & dementia**
- de Hoog et al. 2017 placed in `genescene.net` → **Osler** diristry for **precision medicine**
- Sansaniwal & Kumar 2015 placed in `brainwatch.net` → **Gaia** diristry for **green tech & ecology** resources

# Conclusion

- Targeted evaluation of FAIR Metrics by humans allows systematic comparison of pairs of papers (each pair with test and comparison).
- Results in a well-organized document that can serve as substrate for peer review of the peer review.
- Collections of claims and equivalence relationships can guide development of formal ontologies.
- These semantically formatted manual evaluation records using the PDP-DREAM Ontology will provide an annotated data set against which to validate future AI/automated approaches.

# More about the FAIR Metrics


- Craig, A., & Taswell, C. (2018, November). The FAIR metrics of adherence to citation best practices. In *Proceedings of ASIS&T 81st Annual Meeting SIGMET Workshop*. ASIS&T.
- Craig, A., & Taswell, C. (2018, December). Formulation of FAIR metrics for primary research articles. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1632-1635). IEEE.
- Craig, A., Ambati, A., Dutta, S., Mehrotra, A., Taswell, S. K., & Taswell, C. (2019). Definitions, formulas, and simulated examples for plagiarism detection with FAIR metrics. In *Proceedings of the Association for Information Science and Technology*, 56(1), 51-57. ASIS&T.

# More about the FAIR Metrics Continued


- Craig, A., Ambati, A., Dutta, S., Kowshik, P., Nori, S., Taswell, S. K., Wu, Q. & Taswell, C. (2019, June). DREAM Principles and FAIR Metrics from the PORTAL-DOORS Project for the Semantic Web 2019. In *IEEE 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1-10. IEEE.
- Craig, A., Athreya, A., & Taswell, C. (2023, October). Example evaluations of plagiarism cases using FAIR Metrics and the PDP-DREAM Ontology. In *2023 IEEE 19th International Conference on e-Science (e-Science)* (pp. 1-2). IEEE.

# More about the PORTAL-DOORS Project


- Taswell, C. (2007). DOORS to the semantic web and grid with a PORTAL for biomedical computing. In *IEEE Transactions on Information Technology in Biomedicine*, 12(2), 191-204. IEEE.
- Taswell, C. (2010). A distributed infrastructure for metadata about metadata: The HDMM architectural style and PORTAL-DOORS system. In *Future Internet*, 2(2), 156-189. MDPI.
- Craig, A. & Taswell, C. (2021). PDP-DREAM Software for Integrating Multimedia Data with Interoperable Repositories. In *Brainiacs*, Volume 2 Issue 1 Edoc HA46280EF, also available via DOI 10.48085/HA46280EF. BHA.

# More about the DREAM Principles


- Taswell, S. K., Triggle, C., Vayo, J., Dutta, S., & Taswell, C. (2020). The hitchhiker's guide to scholarly research integrity. In *Proceedings of the Association for Information Science and Technology*, 57(1), e223.
- Athreya, A., Taswell, S. K., Mashkoor, S., & Taswell, C. (2020, September). Essential question: 'equal or equivalent entities?' about two things as same, similar, or different. In *2020 Second International Conference on Transdisciplinary AI (TransAI)* (pp. 123-124). IEEE.
- Craig, A., Lee, C., Bala, N. & Taswell, C. (2022). Motivating and Maintaining Ethics, Equity, Effectiveness, Efficiency, and Expertise in Peer Review. *Brainiacs Journal of Brain Imaging And Computing Sciences*, 2022, 3.

# Contact Info


- acraig@bhavi.us
- www.BHAVI.us
- www.BrainHealthAlliance.org

# References



Foster, Evangeline M, Adrià Dangla-Valls, et al. (2019). "Clusterin in Alzheimer's disease: mechanisms, genetics, and lessons from other pathologies". In: *Frontiers in Neuroscience* 13, 164. eid:164, doi:10.3389/fnins.2019.00164. DOI: 10.3389/fnins.2019.00164.


Gnat, Sebastian et al. (2022). "Retraction of: Sebastian Gnat, Aneta Nowakiewicz, Przemysław Zięba: Taxonomy of dermatophytes". In: *Adv Microbiol* 61.4. doi:10.2478/am-2022-013, pp. 261–261. DOI: doi:10.2478/am-2022-013.


Hoog, G Sybren de, Karolina Dukik, et al. (2017). "Toward a novel multilocus phylogenetic taxonomy for the dermatophytes". In: *Mycopathologia* 182, pp. 5–31.


Mons, Barend (2005). "Which gene did you mean?" In: *BMC Bioinformatics* 6, 142. eid:142, doi:10.1186/1471-2105-6-142. DOI: 10.1186/1471-2105-6-142.


Sansaniwal, SK and M Kumar (2015). "Analysis of ginger drying". In: *Journal of Mechanical Engineering and Sciences* 9. doi:10.15282/jmes.9.2015.13.0161, pp. 1671–1685. DOI: 10.15282/jmes.9.2015.13.0161.


Taswell, Carl (Mar. 2007). "DOORS to the Semantic Web and Grid with a PORTAL for Biomedical Computing". In: *IEEE Transactions on Information Technology in Biomedicine* 12.2 (2). In the Special Section on Bio-Grid published online 3 Aug. 2007, pp. 191–204. ISSN: 1089-7771. DOI: 10.1109/TITB.2007.905861.


Uddin, Sahab, Tanvir Kabir, et al. (2022). "Retraction Note to: Exploring the Role of CLU in the Pathogenesis of Alzheimer's Disease". In: *Neurotoxicity Research* 40.4. doi:10.1007/s12640-022-00519-1, pp. 1125–1125. DOI: doi:10.1007/s12640-022-00519-1.


Ullah, Fahim, Min Kang, et al. (2018). "Retracted: Experimentally investigated the asparagus (Asparagus officinalis)". In: *Food Science & Nutrition* 6.6, pp. 1357–1357.


Wilkinson, Mark, Michel Dumontier, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3, 160018. eid:160018, doi:10.1038/sdata.2016.18. DOI: 10.1038/sdata.2016.18.