Ethical Issues in Language Testing Research

JALT TEVAL SIG Webinar

Daniel R. Isbell
disbell@hawaii.edu

University of Hawai'i at Mānoa
Department of Second Language Studies | Ke ke'ena a'o 'ōlelo hou

Overview

1. Research Ethics – A Broader View
2. Ethics in Language Testing Research
3. Promoting Research Ethics in Language Testing
4. Q&A

Warm-up

What is the first thing that comes to mind when you think of “research ethics”?

Type your answer into the chat.

Research Ethics

Often..

- Focused on institutional, ‘macro-ethical’ issues (Kubanyiova, 2008)
  - Human subject protections
    - Informed consent
    - Privacy
  - Study approval by a review board
- Receives little focus in training for applied linguists
  - Limited coverage in syllabi and textbooks (Wood et al., 2024)

Research Ethics – A Broader View

- Protection of research participants
- Mentor and mentee responsibilities
- Peer review
- Data acquisition, management, sharing, and ownership
- Publication practices and authorship
- Conflicts of interest
- Research misconduct

Steneck (2007)

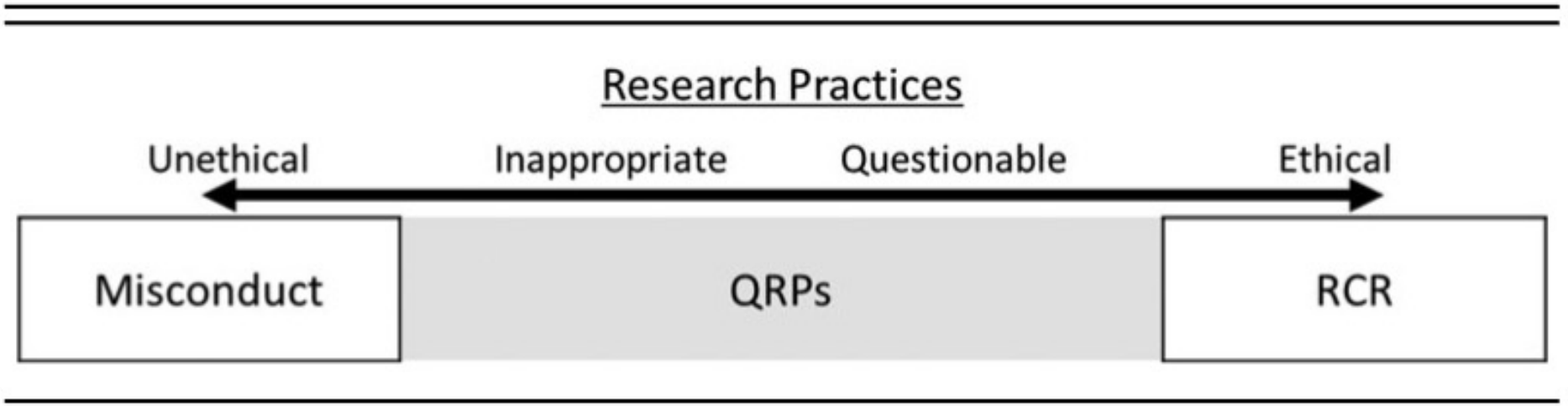
Research Ethics – A Broader View

- Protection of research participants
- Mentor and mentee responsibilities
- Peer review
- **Data acquisition, management, sharing, and ownership**
- **Publication practices and authorship**
- **Conflicts of interest**
- **Research misconduct**

Steneck (2007)

Ethical Research Practices - Key Concepts

- Responsible Conduct of Research (RCR)
- Research Misconduct
- Questionable Research Practices (QRPs)



Responsible Conduct of Research

- RCR: how research should be done (from start to finish)
  - Integrity: Following ‘best practices’ of the research community to pursue truth/generate knowledge
  - Ethics: Application of moral principles to research

- “Study quality as an… ethical imperative” (Plonsky, 2024)
  - Methodologically rigorous
  - Transparent
  - Ethical
  - Valuable to society (and not wasteful, see Isaacs & Chalmers, 2023)

Research Misconduct

- Fraud: Making up data and/or results
- Falsification: Manipulating materials, equipment, data, or results to distort findings

^ these are more common than we'd like to see in AppLing (~17%, Isbell et al, 2022)

- Plagiarism: Copying or otherwise taking credit for the work of others

Deliberate Actions

Questionable Research Practices

- Not (quite) misconduct and often without bad intention
- Include things like…
  - Inappropriate rounding of *p* values
  - Omitting results that do not favor your hypotheses
  - Omitting methodological details/complications to make a study seem more polished
  - Not reporting expected statistical information (e.g., SDs, effect sizes)
  - See Isbell et al. (2022) and Larsson et al. (2023) for more.
- Very, very common in applied linguistics research (> 90% of us have done these at some point)

Questionable Research Practices

Why do QRPs happen?

- Honest Mistakes
- Sloppiness
- Ignorance
- Genuine lack of consensus in a field
- Motivated reasoning
  - Different from outright dishonesty;
  - E.g., not looking for flaws when something works out the way you'd hope at first glance

Conflicts of Interest

- Conflict of Interest (COI) is a pillar of research ethics (Steneck, 2006)

  (yes, right alongside ethical treatment of human & animal participants)
- COIs can influence research agenda setting, analytical decisions, reporting
  - “file drawer problem” (Rosenthal, 1979)
  - “garden of forking paths” (Gelman & Loken, 2013)
  - Fraud & questionable research practices (Isbell et al., 2022)
- Stakeholders should have access to trustworthy information about tests

Why and how some studies are published:

The Uber files

Uber paid academics six-figure sums for research to feed to the media

High-profile professors in Europe and the US were engaged as part of lobbying campaign, leak shows

Felicity Lawrence

Tue 12 Jul 2022 01.00 EDT

RESEARCH | ACADEMIC CAPTURE | ANTITRUST AND COMPETITION | CORPORATE GOVERNANCE | REGULATORY CAPTURE | RENT SEEKING

Uber and the Sherlock Holmes Principle: How Control of Data Can Lead to Biased Academic Research

BY LUIGI ZINGALES *October 9, 2019*

And why other studies are not:

Big oil and gas kept a dirty secret for decades. Now they may pay the price

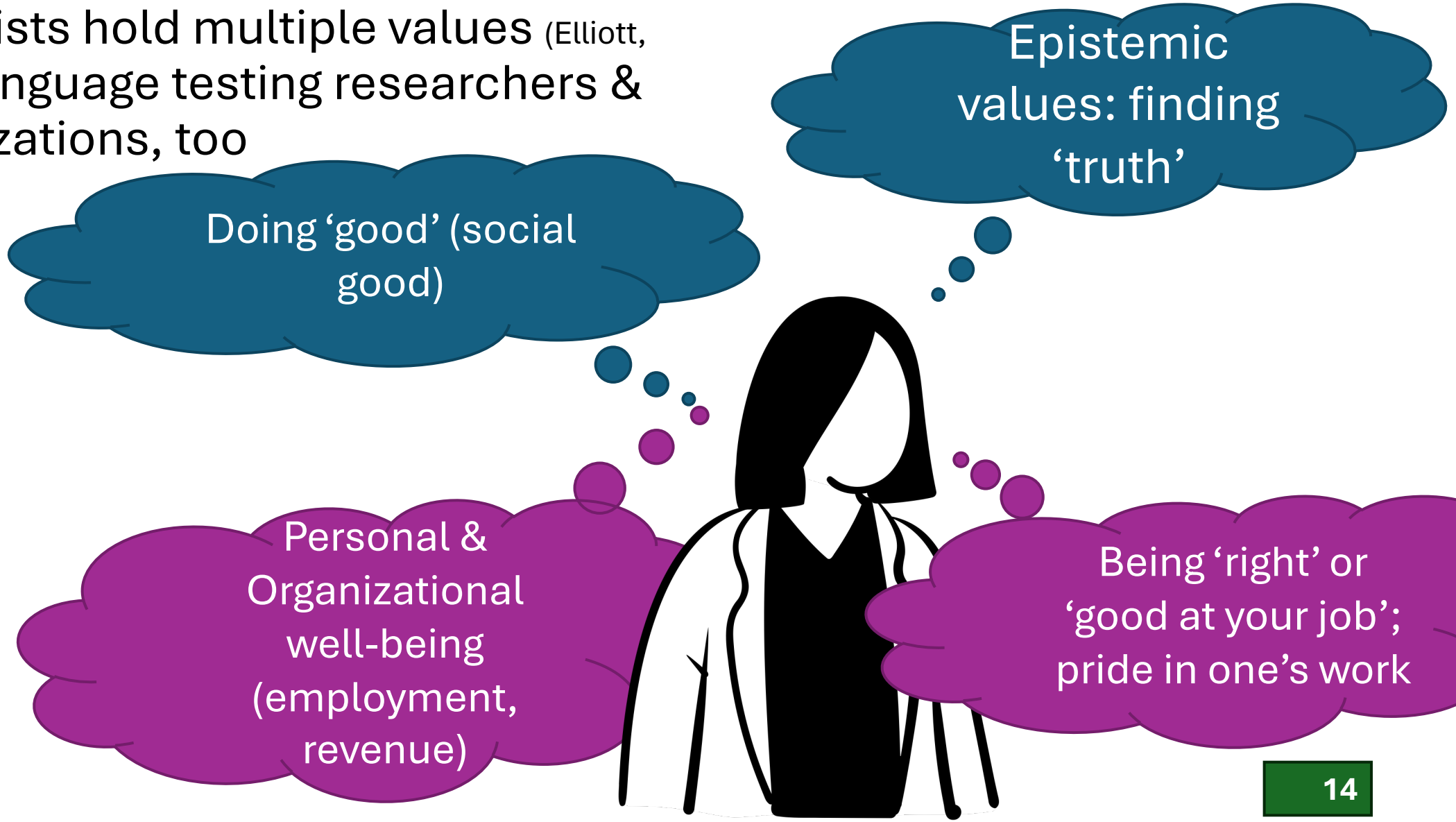
Chris McGreal

Wed 30 Jun 2021 03.00 EDT

But, even more strikingly, the nearly two dozen lawsuits are underpinned by accusations that the industry severely aggravated the environmental crisis with a decades-long campaign of lies and deceit to suppress warnings from their own scientists about the impact of fossil fuels on the climate and dupe the American public.

Tensions among Values

Scientists hold multiple values (Elliott, 2022); language testing researchers & organizations, too



Transparency: Statements of (Potential) COIs

- Stakeholders rely on (peer-reviewed) research to make informed decisions about policy
  - Developer publications (websites, white papers) transparent but may be perceived as less objective/trustworthy
  - Peer-reviewed research is assumed to be more rigorous and vetted; ‘gold standard’
- Transparency and disclosure is seen as the best way to handle **potential** COIs
  - COI / “Competing Interest” statements are not admissions of guilt
  - Help readers judge sources with less ambiguity

Ethics in Language Testing & Language Testing Research

Ethical Guidance in Language Testing

- ILTA Guidelines for Practice (2018-2022): https://www.iltaonline.com/page/ILTAGuidelinesforPractice
- ILTA Code of Ethics (2000/2018): https://www.iltaonline.com/general/custom.asp?page=CodeofEthics

Ethical Guidance in Language Testing

- These documents are excellent, but focus mostly on language testing *practice*
  - Development and administration of tests, use of test scores, etc.
  - Rendering professional services to governments and organizations (advising/consulting)
- Somewhat limited (but not absent!) guidance for doing research

ILTA Code of Ethics, Principle 3

"Language testers should adhere to all relevant ethical principles embodied in national and international guidelines when undertaking any trial, experiment, treatment or other research activity."

ILTA Code of Ethics, Principle 3, Annotations

- Language testing progress depends on research, which necessarily involves the participation of human subjects. This research shall conform to generally accepted principles of academic inquiry, be based on a thorough knowledge of the professional literature; and **be planned and executed according to the highest standards**.
- **All research must be justified**; that is proposed studies shall be reasonably expected to provide answers to questions posed.
- The **human rights of the research subject** shall always take precedence over the interests of science or society.

ILTA Code of Ethics, Principle 3, Annotations

- Where there are likely discomforts or risks to the research subject, the benefits of that research should be taken into account but must not be used in themselves to justify such discomforts or risks. If unforeseeable harmful effects occur, the research should always be stopped or modified.
- An **independent Ethics Committee should evaluate all research** proposals in order to ensure that studies conform to the highest scientific and ethical standards.
- Relevant information about the aims, methods, risks and discomforts of the research shall be given to the subject in advance. The information shall be conveyed in such a way that it is fully understood. Consent shall be free, without pressure, coercion or duress.

ILTA Code of Ethics, Principle 3, Annotations

- The subject shall be **free to refuse to participate in or to withdraw** from, the research at any time prior to publication of research results. Such refusal shall not jeopardise the subject's treatment.
- Special care shall be taken with regard to obtaining prior consent in the case of subjects who are in dependent relationships (for example, students, the elderly, proficiency challenged learners).
- In the case of a minor, consent shall be obtained from a parent or guardian but also from the child if he is of sufficient maturity and understanding.

ILTA Code of Ethics, Principle 3, Annotations

- Confidential information obtained in research shall not be used for purposes other than those specified in the approved research protocol.
- Publication of research results shall be **truthful and accurate**.
- Publication of research reports **shall not permit identification** of the subjects who have been involved.

My Commentary

- Primary focus on **institutional ethics** (macro-ethics, in Kubanyiova's 2008 terms)
  - Protection and rights of participants
  - Consent
- Other aspects of research ethics present, but not specified (defers to other sources/standards)
  - Study justification (flipside: research waste)
  - Rigor and Appropriateness of analyses
  - Accuracy of reporting
- Some aspects missing
  - Conflicts of interest, transparency & reproducibility/reusability

Some Problems We (May) Face in Language Testing Research

Maybe: Misconduct and QRPs

- Language testers are highly concerned with the reliability of test scores and the validity of test score use
- Are they as concerned with the validity and rigor of research (Shadish et al., 2002)?
  - Construct Validity
  - Internal Validity
  - External Validity
  - Statistical Conclusions Validity

Construct Validity

- Relevant to how well measured (or manipulated) variables relate to the theoretical construct(s) of interest
- We're probably fine with this!
- At least when the constructs of interest are related to language knowledge/skills/abilities
- Related to QRPs:
  - Do we choose easy/convenient instruments, or ones which are thought to best capture the constructs of interest?

Internal Validity

- Relevant to causal relationships among variables
- A potential challenge, and probably lots of variability
- A lot of LT research is observational, which creates challenges for causal interpretations
- How good are experimental designs in LT? Observational designs?
- Related to QRPs:
  - Do we adopt the strongest study designs? (experimental designs, randomization, etc.)
  - Is our research adequately powered?

External Validity

- Relevant to generalizability of findings
- Mixed bag:
  - Some studies in LT very large and representative thanks to operational test data
  - Others have smaller, convenience samples
- We are probably not clear enough about our sampling and choices in research design/analytical decisions
- Few replications that can help understand the generalizability of findings
  - Some meta-analyses, which are very helpful – more would be nice

Statistical Conclusions Validity

- Language testers have a reputation for being good at stats, but…
- LT research is not necessarily more rigorous or better reported than other areas in Applied Linguistics
  - Not in bad shape, but other areas have advanced
- Aryadoust et al. (2021) found that reporting on Rasch model assumptions had several shortcomings
  - Item separation
  - Unidimentionality
  - Local item dependence
- QRPs: Do we take ‘shortcuts’ or take some analyses for granted and fail to check assumptions and report results fully?

Definitely: Conflicts of Interest

- Conflict of Interest (COI) is a pillar of research ethics (Steneck, 2006)

  (yes, right alongside ethical treatment of human & animal participants)
- COIs can influence research agenda setting, analytical decisions, reporting
  - “file drawer problem” (Rosenthal, 1979)
  - “garden of forking paths” (Gelman & Loken, 2013)
  - Fraud & questionable research practices (Isbell et al., 2022)
- Stakeholders should have access to trustworthy information about tests

Testing Research and Public Reasoning

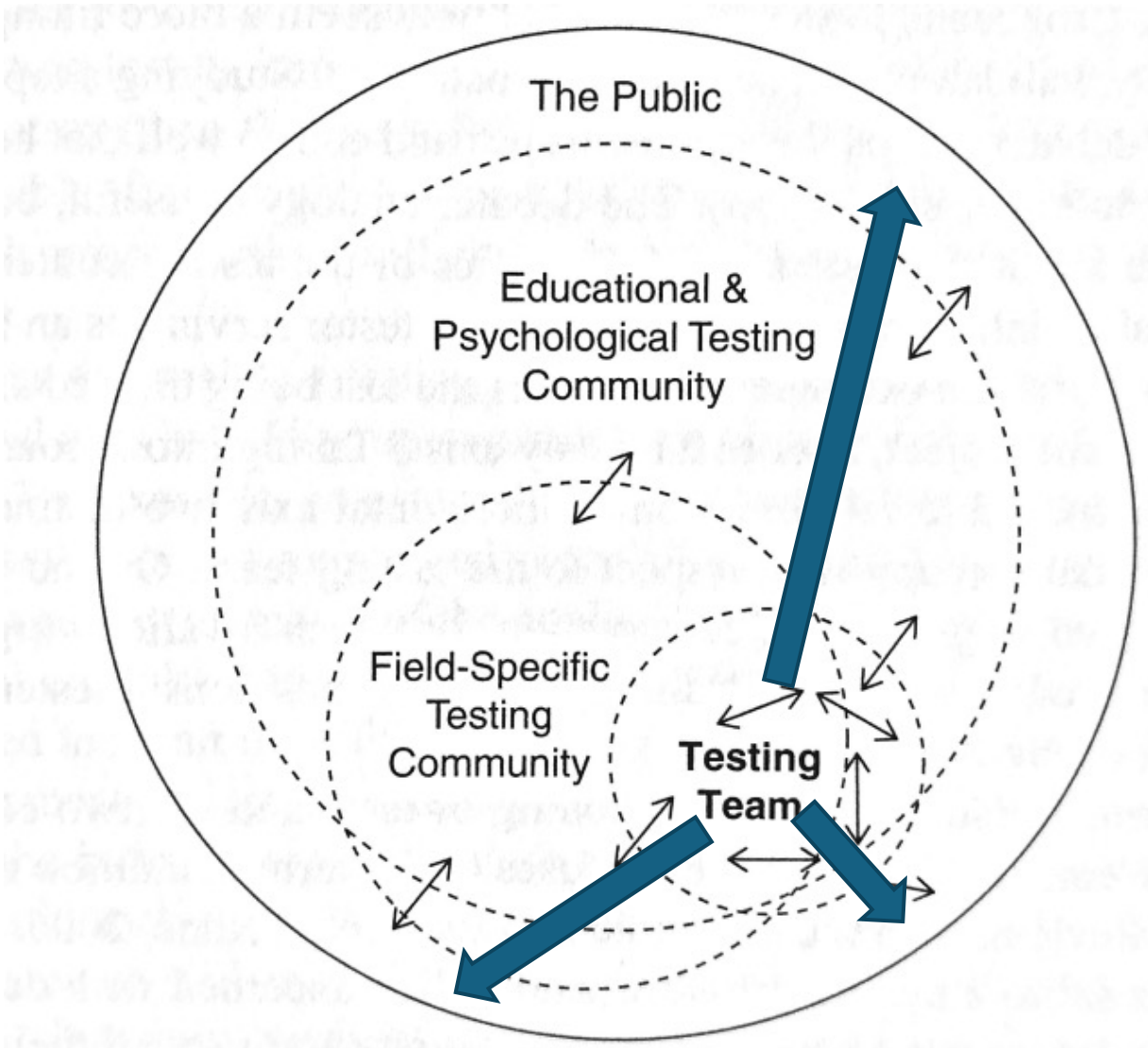


Figure 7.2 from Chapelle (2021, p. 107)

Independence in Test Validation Research

"Once the test and the IUA [Interpretation/Use Argument] are developed, the focus shifts (especially for high stakes applications) and **a more critical and arm's length evaluation of the proposed interpretation and use** can be adopted. In the appraisal stage, **the IUA should be challenged, preferably by a neutral or skeptical evaluator**. If the validity of the proposed IUA is to be evaluated by the assessment developers, as is often the case, they should seek to identify and examine the challenges that might be posed by a skeptical critic."

(Kane, 2013, p. 17, emphases added)

The “Who” of Validation Research

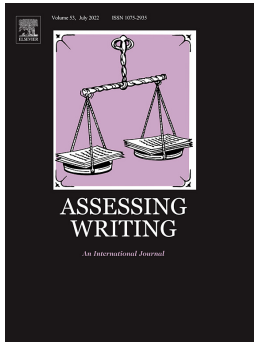
Who sets the validation agenda?

Who selects/designs validation studies?

Who enables validation studies?

Who conducts and reports validation studies?

COI statement or beating around the bush?



Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects

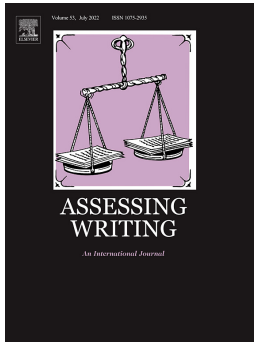
Daniel R. Isbell

Michigan State University, B331 Wells Hall, 619 Red Cedar Road, East Lansing, MI 48824, USA

Acknowledgements

This study was made possible by the support of English Language Center Testing Office at Michigan State University, which provided access to test score data and test materials. The author wishes to thank Dr. Ryan Bowles for statistical advice and comments

COI statement or beating around the bush?



Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects

Daniel R. Isbell

Michigan State University, B331 Wells Hall, 619 Red Cedar Road, East Lansing, MI 48824, USA

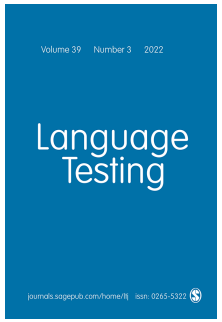
I was working as an RA in this office.

Acknowledgements

This study was made possible by the support of English Language Center Testing Office at Michigan State University, which provided access to test score data and test materials. The author wishes to thank Dr. Ryan Bowles for statistical advice and comments

No distinct, explicit COI statement

Conflicting COIs?



[redacted]

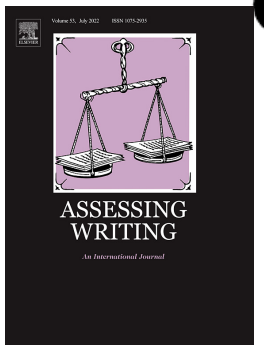
Educational Testing Service, USA

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The research reported in this paper was funded by the *TOEIC* program at Educational Testing Service. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.



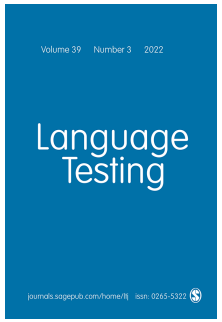
[redacted]

Educational Testing Service, USA

Acknowledgements

This research was funded by the TOEIC program, and we are employed by Educational Testing Service, who has an ownership stake in the TOEIC test.

Conflicting COIs?



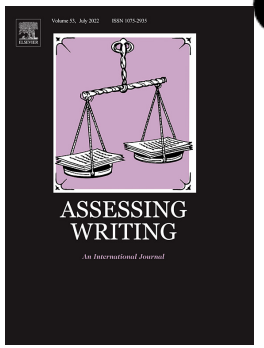
[redacted]
Educational Testing Service, USA

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The research reported in this paper was funded by the *TOEIC* program at Educational Testing Service. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.



[redacted]
Educational Testing Service, USA

Acknowledgements

This research was funded by the TOEIC program, and we are employed by Educational Testing Service, who has an ownership stake in the TOEIC test.

COI Disclosure

Studies with a conflict disclosure:

3

2% of all 181 studies,
3% of all studies 117 with a COI statement

COI Disclosure

Studies with a conflict disclosure:

3

2% of all studies,
3% of all studies with a COI statement

"This research was funded by the TOEIC program, and we are employed by Educational Testing Service, who has an ownership stake in the TOEIC Test." (Schmidgall & Powers, 2020, p. 12)

"The main author is not employed by Trinity, the co-author is employed by Trinity." (Harsch & Kanistra, 2020, p. 281)

"KN is employed by the Eiken Foundation of Japan, which develops and administers the TEAP test." (In'nami et al., 2016, p. 22)

How many articles **should** have a disclosure?

Studies with a conflict disclosure:

3

2% of all studies

Studies with an authorship conflict:

46

25% of all studies

How many articles **should** have a disclosure?

Studies with a conflict disclosure:

3

2% of all studies

+ Previous affiliations, membership on advisory committees, etc.

-Could be *underestimating*

Studies with an authorship or other potential conflict:

67

37% of all studies

In other words:

COI are substantially underreported (4-7% of needed disclosures)

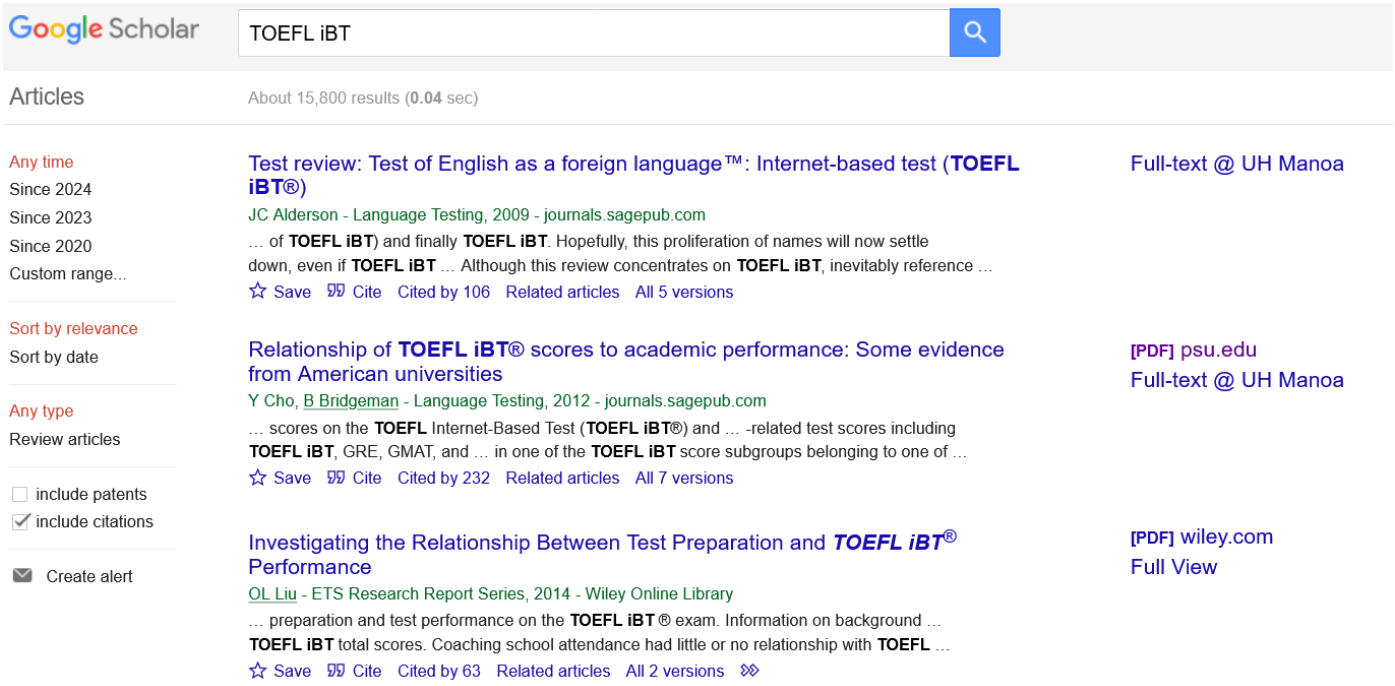
Many published COI ("Competing interests") statements are inaccurate

- *pro forma* exercise

Disclaimers appear more commonly than COI disclosures

"Any opinions expressed in this paper are those of the authors and not necessarily of TEST DEVELOPER"
– report by researchers all employed by TEST DEVELOPER

Activity:



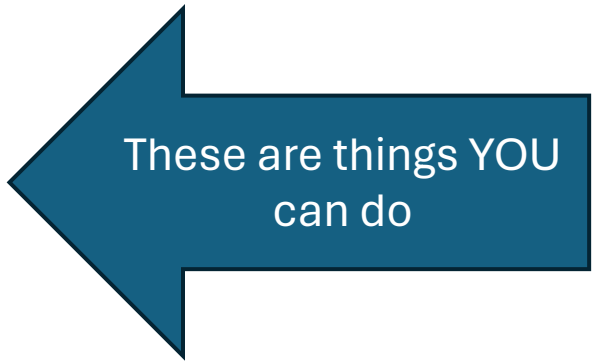
1. Type the name of a well-known test into Google Scholar
2. Click on some of the links (don't need access to the full article)
3. Who is conducting the studies? Who do they work for?
4. Can you find a COI statement? Acknowledgment of funding?

Promoting Research Ethics in Language Testing

Promoting Research Ethics in Language Testing

- Review boards (where applicable) and strong concern for and actions to protect participants: YES… and…

- Transparency:
  - Funding acknowledgements
  - Conflict of interest disclosure (Isbell & Kim, 2023)
  - Open Science practices (Winke, forthcoming)



Disclose your COIs

- Focus on current and recent associations (last 5 years)
- Not limited to financial stakes in a test
- “What might look like a source of bias to a reasonable person?”
  - You work for the developer of the test or an affiliate
  - You consult for the developer of the test
  - You were the one who designed the test

Some recent examples at *Language Testing*

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Ruslan Suvorov currently serves as Associate Editor of Language Testing. He was blinded to the manuscript in the ScholarOne online submission platform and Dr. Talia Isaacs managed all stages of its processing as handling editor. The remaining co-authors declared no potential conflicts of interest with respect to the research and authorship of this study.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The first author (Neittaanmäki) is employed as a research statistician for the National Certificates of Language Proficiency (NCLP) examination system. However, she was not involved in planning and drafting tasks nor assessing performances. NCLP is administered by the Finnish National Agency for Education, funded by the Ministry of Education and Culture and operated by the University of Jyväskylä. NCLP researchers are complementarily financed by the University of Jyväskylä.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: During the last five years, the first author, Ute Knoch, conducted assessment-related research or consultancy work for the following organisations: Educational Testing Service (ETS), IELTS, Pearson, Cambridge Boxhill Language Assessments, Australian Department of Defense, Australian Civil Aviation Safety Authority, Australian Health Practitioner Regulation Authority, Benesse Corporation, Australian Department of Home Affairs. She served, until 2021, on the Pearson Technical Advisory Board and is the current test review editor of Language Testing. The second author, Jason Fan, conducted assessment-related research, advisory, or consultancy work for the following organisations: British Council, Pearson Education, PeopleCert, Cambridge Boxhill Language Assessment, Educational Testing Service, and Language Training and Testing Centre (LTTC).

Engage in Open Science

- More than just Open Access publishing (which is great, by the way!)
- Share your…
  - Study data
  - Study materials
  - Analysis codes/scripts
- Preregister your study

Not always easy to do, but worthwhile
(Al-Hoorie et al., 2023; Winke, forthcoming)

Data Sharing

- Good: Share your processed/clean data used in your reported analyses
- Better: Also share your ‘raw’ data
- Best: Include a “data dictionary” that helps others understand your data
- Always:
  - use non-proprietary formats when possible (.csv files, not SPSS .sav files)
  - Share your data on a publicly-accessible repository (unless it is too sensitive to share in that way)
    - Use Open Science Framework (osf.io) and not a journal’s hosting service

Materials Sharing

- LT does pretty well at this, comparatively, when it comes to test materials
  - Mainly because of practice/mock tests that are publicly available
- Share your other instruments when possible
  - Surveys/questionnaires
  - Test materials or other tasks/experiments/etc. created for research purposes
- Share on a publicly-accessible repository (e.g., OSF, iris-database.org)

Code/Analysis Script Sharing

- This potentially lets other researchers follow your exact steps
- For quant research:
  - R scripts, JASP/JAMOVI scripts, SPSS scripts
  - Try to use comments and write code clearly so others can ‘read’ it
    - No one is expecting you to be a professional software engineer; don’t worry about people judging your code as long as it works and can reproduce your results
- For qual research:
  - Coding schemes/categories
  - Matrices you used to organize/refine your coding
  - Computer files from CAQDAS programs

Preregistration

- Definition: Creating and registering a research plan (study design) before you start the study
  - Includes RQs, hypotheses, and full methodological details
  - Timestamped
- Does not have to be a *registered report*
- You can choose when to share the preregistration publicly, but should be shared with editors/reviewers when you submit the article
  - OSF is again a good option for this

In the end…

- A bit of extra work, but some simple things we can do to make our language testing research more ethical through transparency
- Benefits to:
  - Public accountability
  - Peer review process
  - Community of language testing researchers

Thank you! Mahalo!
ありがとうございました！

Questions? Comments?

disbell@hawaii.edu

https://www.hawaii.edu/sls

What is an ethical dilemma you've faced in language testing research?

Have you seen any ethically questionable language testing research? What about it made you uncomfortable/skeptical?