UNIVERSITY OF CAMBRIDGE

From the outside in

Fairness in assessment – looking forward from the pandemic

Isabel Nisbet

IAEA Conference, 7 October 2021

Faculty of Education

Fundamental questions for the assessment community

- Is what we do fair?
- Do those whom we serve think that what we do is fair?
- What can we do about it?

Outline

- What is fairness and why does it matter?
- Challenges to fair assessment: theoretical challenges
- Challenges from critical theory and culture wars
- Challenges from the Covid experience
- Fair assessment – from the outside in
- Questions to take away

Why does fair assessment matter?

- Fairness is a fundamental moral/professional value
- Increasing controversy about the fairness of national/state exams and tests
- Worry about the use of tests for college/university entry (“test-optional” approaches in the USA)
- Debates about fair selection for employment (“Uniform Guidelines” (USA) – the “diversity/validity dilemma” – tests as thermometers?)
- Cheating – epitome of unfairness
- Increased academic focus on “fairness” in assessment (the “big two” become the “big three”) (Worrell, 2016)

And ask any parent ……

- Concern about unfairness is a “pre-social” emotion, starting in infancy and continuing throughout life

… and more recently

- The Covid experience
  - Existing inequalities and injustices in society widened
  - Attempts to assess fairly scrutinised and criticised
- Controversy/debate/division about values in society: ever-present and increasingly vocal
  - Clusters of conflicting ideologies affecting education and assessment (see Putnam “The Upswing” (2020) - also “The reading wars” (Pearson, 2004), “Race, retrenchment and the reform of school mathematics” (Tait, 1994))

Example: Issues of racial inequality (BLM) at an assessment conference in Spring 2021

What does “fairness” mean?

Several distinct senses

1) (purely formal) accurate/appropriate (“a fair copy”)
2) (implied contractual) meeting legitimate expectations of those affected
3) (relational) treating like cases alike – a sort of equality
   - in relevant respects (but what is relevant?)
4) (retributive) an appropriate reward (or penalty) for prior behaviour. The outcome is deserved
5) (consequential) leading to fair/socially just outcomes
6) (differential esteem) violated by disrespect (or causing legitimate offence) to a sub-group

What does “fairness” mean?

Several distinct senses

1) (purely formal) accurate/appropriate (“a fair copy”)
2) (implied contractual) meeting legitimate expectations of those affected
3) (relational) treating like cases alike – a sort of equality
   - in relevant respects (but what is relevant?)
4) (retributive) an appropriate reward (or penalty) for prior behaviour
5) (consequential) leading to fair/socially just outcomes
6) (differential esteem) violated by disrespect (or causing legitimate offence) to a sub-group

Why these definitions matter – a Scottish philosophical joke ☹

Theoretical challenges: problems with the received view of assessment fairness

- The received view: fairness/unfairness is to candidates/test-takers (or groups of candidates)
- Unfairness happens when outcomes for different [groups of] candidates vary because of construct-irrelevant attributes (e.g. social class, gender, disability)
- Good practice is to minimise risk of unfairness in advance at the design stage (universal design) and use analysis in arrears to investigate possible unfairness (Differential [item] functioning analysis)
- BUT focus on relational fairness without asking why it matters
- Sense of (un)fairness is much weaker if no competitive access to limited goods
- Too much focus on groups

“Conditional fairness” (Mislevy, 2013, 2018)

- Prompted by increasing awareness and understanding of diversity among test-takers
- “Making tests identical for all examinees can make them less fair”
- The same content may not be equally meaningful/accessible to all candidates
- Need to put the “assessment argument” (tasks/evidence/inference) in context: identical assessment tasks may not provide the same information about the knowledge and skills of all candidates
- These are not (statistically) random differences, but should be taken into account from the outset in the assessment argument
- Theoretical structure for “universal design” taking account of differences from the outset, not just “retro-fitting” a standard design where needed

My response to Mislevy's arguments

- Insight that the assessment argument (tasks – evidence - inferences) should be seen in context
- Agree that the principles of universal design can support fairer assessment
- Agree that varied tasks, informed by universal design, can lead to equivalent outcomes
- **But is equivalence [comparability] always required for fairness**? What if it is not essential for a fair outcome? (eg applications for local colleges/universities)?

Fairness and validity

- What is the relation between validity and fairness?
- My previous view (Nisbet & Shaw 2020): Validity is a necessary but not sufficient condition for fairness
- My developing view (building on others, eg Stobart): 3 approaches to validity:
    - Internal (coherence of the structure and content of the assessment for its intended purpose)
    - Marginal: takes into account links with selected real-life contextual factors before the test (eg opportunity to learn) and after the test (validity of use)
    - External (situated): starts with the assessment as a real-world event in a real context and derives conclusions about the assessment

Fairness and validity

- What is the relation between validity and fairness?
- My previous view (Nisbet & Shaw 2020): Validity is a necessary but not sufficient condition for fairness
- My developing view:
  - 3 approaches to validity:
    - Internal (coherence of the structure and content of the assessment for its intended purpose)
    - Marginal: take into account links with some real-life contextual factors before the test (eg opportunity to learn) and after (validity of use)
    - External: starts with the assessment as a real-world event in a real context and derives conclusions about the assessment

Fairness and validity

- Assessment theorists have seen fairness as a [now adult] offspring of validity: starting with validity and developing ideas of fairness from it
  - [Received view of fairness has its origins in a well-documented, narrow, conceptualisation of validity]
- The outside-in approach: starts with fairness as a societal concept and applies it to activities, including assessment (also educational policy choices, curriculum design….)

Challenges from critical theory

- Many variants – critical race theory, critical discourse analysis, critical disability studies, institutional ethnography…….
- Key messages: institutions – including schools, assessment organisations, voluntary and imposed groupings of individuals – can be structurally unjust/biased and their activities can reflect/reinforce the underlying power relationships
- Intersectionality: complex interrelationship of social categories (eg gender, ethnicity, religion, poverty, health..); underlying power relationships need to be understood (see, eg, MW Apple, 2019) – “the poor black unhealthy single mother”
- Link to questions of fair assessment – we need to stand back and consider the institutional context and the underlying power relationships. Who benefits from our assessments? Who does not benefit?

My learning from the challenges of critical theory

- Critical approaches ask legitimate questions, starting from what is right for society
- It's right for us to look critically at the institutions we work in and realise that we are influenced by them
- Psychometric approaches (including DIF analysis) can break down their material into categories that fail to take account of intersectionality
    - What if the categories/comparators used for DIF analysis are themselves biased?
    - What if all assessments of some kind have always been unfair? (is reliability enough? What about consistent/repeatable unfairness?)

But…

- Critical approaches are **not alternatives to good micro-practice** to improve the fairness of assessment (eg universal design)- there is always a place for “disciplined deep dives” into good practice (MW Apple)

Challenges to assessment from the Covid experience

- Need to adapt/improvise – normal arrangements (eg for attending for examinations) sometimes not possible
- Timescale for changes was faster than normal in the sector – modelling and trialling was difficult/impossible
- Institutions – and the public - had to overcome their natural caution about the use of technology in assessment
- Attempts in some countries to estimate marks by statistical methods - using an “algorithm” - were badly received

Example (UK): the ‘mutant’ algorithm (2020)

- Decision that it was impossible to hold (mainly terminal, written) exams in the normal way
- Aim: relational fairness across the cohort as a whole – “fairness across a whole population” (Ofqual)
- Statistical models developed to predict grades based on historical information about schools and individuals and rank orderings by teachers
- Models consulted on and extensively tested
- But outcry when they were used – U-turn by governments – grades proposed by teachers awarded instead – significant grade inflation – unfairnesses between types of school/subjects

What's unfair about algorithms?

- The aim was for relational fairness, but the approach failed to achieve the second wing of fairness – desert. Students/parents felt that the grades did not reflect what they had done
- The use of statistics can help to achieve relational fairness across a cohort as a whole
- But relational fairness in the aggregate must be balanced with individual desert
- Algorithms work better with human engagement
- Humans work better with algorithmic supervision – emerging findings about possible bias in teacher judgement (gender/race)

The algorithm v teacher assessment - Underlying fairness problem

- A teacher thinks that 10 of her students have a strong probability (say 80%) of getting a grade A
- All 10 students are told that they are “Grade A level”, and are expecting an A
- But in “normal” times, 2 out of 10 would not perform so well on the day in an exam and would get a B
- The algorithm – aiming to make the present like the past and the same for everyone - awards a Grade A to the top 8 in the group (according to the teacher’s rank ordering) and a B to the bottom 2
- The bottom 2 who receive a B are outraged that their grade has been lowered by a “mutant algorithm”.
- **Which is the fairer mark for the bottom 2 – A or B?**

Fair assessment – from the outside in

- My emerging view: Fairness (in assessment) should be seen as a societal concept, not as a specialist offspring from validity
- Both of the component ideas of fairness – equality and desert need to be considered. The assessment community has tended to overlook the importance of individual desert. The revolt against the algorithm brought that out.
- Relational fairness may not always require equivalence/comparability – at a deeper level it may be relationally fair to enable each student to show what they can do (in different ways)
- Reliability may be an irrelevance – the whole system may be unfair

But it’s an unfair world – what can I do?

- Fairness is not a binary concept, but a spectrum – we can aim to make our assessment fairer (“fertile functioning” (Jonathan Wolff))
- There is still an important place for detailed technical work to improve fairness – in assessment design and analysis of outcomes
- We can ask questions about the contexts in which our assessments are taken – including the big questions posed by the critical theorists
- Where we are aware of contextual factors that may affect the fairness of our assessments as events in the real world, we can talk about them – include them in our reports/publications

Outline

- What is fairness and why does it matter?
- Challenges to fair assessment: theoretical challenges
- Challenges from critical theory and culture wars
- Challenges from the Covid experience
- Fair assessment – from the outside in
- Questions to take away

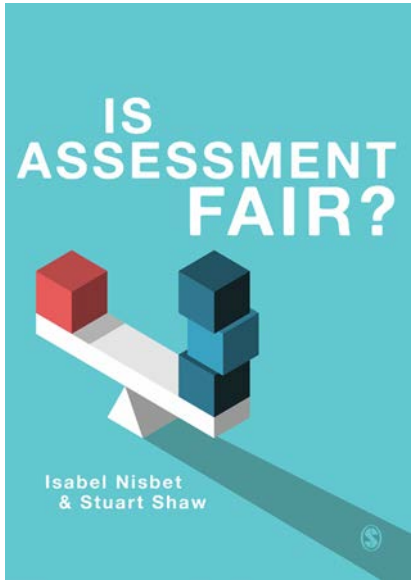
OUTSIDE IN – Questions for the assessment community

- What is the purpose of the assessment that I am working on (“my assessment(s)”)? Who will benefit if this purpose is achieved? Who will not benefit?
- Who have the best opportunities to learn the subject-matter of my assessment? Who have the least opportunities?
- Is my assessment accessible and meaningful to all/most candidates? Any exceptions?
- Is the content/construct/curriculum for my assessment accessible and meaningful to all candidates? Any exceptions?
- (for technical/theoretical work) Are my assumptions/sources/comparators fair?
- In the real world context when my assessment is used, will the outcome be fair? What do I need to know to answer that? WHAT CAN I DO TO MAKE IT FAIRER?

Many thanks

nisbet.isabel@gmail.com

in235@cam.ac.uk



References

- MW APPLE, “On doing critical policy analysis”, *Educational Policy* 2019, Vol. 33(1) 276–287
- Doyin ATEWOLOGUN, “Intersectionality Theory and Practice”, published online 28 August 2018, accessed at https://doi.org/10.1093/acrefore/9780190224851.013.48
- NISBET, I, & SHAW, S. D. (2020). *Is Assessment Fair?* London: SAGE Publishing.
- Robert J. MISLEVY, Geneva Haertel, Britte H. Cheng, Liliana Ructtinger, Angela DeBarger, Elizabeth Murray, David Rose, Jenna Gravel, Alexis M. Colker, Daisy Rutstein & Terry Vendlinski (2013) “A “conditional” sense of fairness in assessment”, *Educational Research and Evaluation*, 19:2-3, 121-140, DOI: 10.1080/13803611.2013.767614
- See also Chapter on A Conditional Sense of Fairness in Mislevy (2018) *Sociocognitive Foundations of Educational Measurement* (Routledge)
- P David PEARSON, “The reading wars”, *Educational Policy*, Vol. 18 No. 1, January and March 2004, 216-252
- Robert D. PUTNAM, with Shaylyn Romney GARRETT, *The Upswing: How we came together a century ago and how we can do it again,* Swift Press, 2020
- William F. TATE, “Race, Retrenchment, and the Reform of School Mathematics” *The Phi Delta Kappan* , Feb., 1994, Vol. 75, No. 6 (Feb., 1994), pp. 477-480, 482- 484