Editorial: Cross-cultural Study of Test effort in PISA

Gavin T. L. Brown & Kane Meissel

The University of Auckland

Correspondence related to this editorial should be addressed to Prof. Gavin Brown, Faculty of Education & Social Work, The University of Auckland, Private Bag 92019, Victoria St. West, Auckland, 1142, New Zealand or by email to gt.brown@auckland.ac.nz

This special issue addresses two challenges. First, the extent to which engagement in international large-scale assessments (ILSAs), specifically the Programme for International Student Assessment (PISA), matter to the scores generated. Second, and more importantly, how test taker engagement can be measured when behavioural data are available.

The OECD uses PISA scores to rank order jurisdictions based on large samples of students within each jurisdiction. Since the 2009 PISA test, the performance of students in Shanghai (2009, 2012), and subsequently the agglomeration of wealthy regions (i.e., Beijing-Shanghai-Jiangsu-Zhejiang) in the People's Republic of China (2018), has been reported to be first in the world. This had substantial impact on politicians and the public in other jurisdictions who compared their own results unfavourably with those of China (Zhao, 2014). Indeed, within the editors' own nation (New Zealand, an OECD member) there have been recent complaints from educational influence groups about our declining scores on PISA, and these are often re-reported in public media (Brown, 2021). Hence, the importance of PISA scores for the perceived academic prowess or standing of a nation and public perception cannot be understated.

The challenge for all ILSAs (including PIRLS, TIMSS) is the level and nature of consequence for the source of the data—the student. It is a *sine qua non* of psychometric theory that score validity depends on test-takers doing their very best throughout the whole test. This assumption is a reasonable expectation when the test has personal consequences (e.g., grade promotion, scholarship, public recognition). But one of the defining characteristics of ILSA systems is that the consequence is not aimed at the student, but rather exists for the system via public dissemination of rank order scores. This means that the level of effort we might expect under the country-reputation stakes associated with PISA might be highly variable across individual students and between different jurisdictions. Research literature shows that this is exactly the case in modern, liberal jurisdictions such as Sweden, that emphasise and reject a high-stakes testing culture. However, we do not yet know if the same conclusion can be reached about students in high-pressure, high-frequency testing cultures, such as China. It is possible that in such contexts, there is no such thing as a low-stakes test on which the test-taker can relax. If that is the case, then comparisons between jurisdictions that have very different ILSA test-taking effort norms are invidious and untenable.

Measuring engagement and effort for any test seems to be currently limited to test-taker self-reports either before or after a test administration, or to monitoring time used per question on a computer-based test. PISA itself uses test-taker self-report via an effort thermometer to estimate the amount of effort students claim to exercise on the PISA test. Other systems of self-reported effort estimation have been developed, but all such methods suffer from potential biases. As with any self-assessment of performance, test-takers may not know very

well how hard they actually tried. They may lie about their effort to please authorities or parents, they may not know what their maximal effort is, they may not know if their effort varied during the test, they may not be able to distinguish objectively knowing from a feeling of knowing, and so on.

To bypass the limitations of self-awareness, response-time effort (RTE) research has used the amount of time spent per question as a proxy for actual engagement or effort, with very low RTE indicating minimal effort. Each test item has a reasonably robust floor for how quickly it can meaningfully be answered, taking into account how much text has to be read and what kind of mental actions are needed to answer the task. Extensive research has shown that students who respond faster than this floor perform poorly and are not making their best effort. Removing such participants improves mean scores. This suggests that, in jurisdictions where lack of effort on low-consequence tests is permissible and extant, rank order scores are relatively false until guessing and low effort are removed. RTE research has shown that effort is variable during a test and suggestions have been made as to how to prompt test-takers to slow down or pay closer attention.

In an effort to extend paper-and-pencil testing from 'vanilla' replication of tasks on a screen, ILSAs are taking advantage of computer-based tests to increase the authenticity of questions. Computer-based testing now offers a wider variety of item formats (e.g., different kinds of multiple choice and constructed responses that can be scored by computer or by a human rater). Interactive problem-solving tasks (Greiff, Wüstenberg, & Avvisati, 2015) generate data not just about *what* answer was chosen but also *how* the test-taker went about arriving at the answer. The computer records a wide variety of actions (e.g., number and order of clicks, time taken on the item as a whole, number of manipulations of task variables, number of times the item is attempted, etc.), all of which can be examined for patterns that may relate to performance and reveal insights about student engagement. Because there are differences in test-taking norms across jurisdictions, the RTE thresholds that we have been using may not be invariant by type of test question and jurisdiction. It may well be that being rapid in different languages on different kinds of items looks quite different, suggesting that universal treatment of data is not warranted.

This special issue

The papers in this special issue provide interesting insights into these two problems. Anran Zhao, Gavin Brown, and Kane Meissel reveal, using self-reported effort for a hypothetical test situation in Shanghai, that the effect of an ILSA test consequence on effort was only marginally less than it would be for a test that had personal consequences. Additionally, this paper also provides validation evidence for the Students' Conceptions of Assessment (Brown, 2008) inventory. Militsa Ivanova, Michalis Michaelides, and Hanna Eklöf revealed that, for Swedish students on constructed response items in the 2015 PISA science test, the number of actions performed correlated with performance, self-reported effort, and location of the item in the test. This suggests a new behavioural indicator for gauging test-taking engagement. Erik Lundgren and Hanna Eklöf examined the behaviours of Swedish students on a PISA 2012 problem-solving task; the multiple actions and time factors indicated that the level of effort invested in the item was not generally related to test-performance, or self-reported test-taking effort. However, clusters of students who put in more effort before giving up on the task had higher test performance, suggesting that RTE by itself does not satisfactorily

distinguish students with low motivation from low effort driven by proper motivation. Hongwen Guo and Kadriye Ercikan examined RTE on PISA 2018 science data across nine jurisdictions representing six languages. Three different methods for estimating RTE were deployed across four different item formats and revealed in general that RTE depended on item difficulty within each language and cultural group, although the magnitude of RTE impact on scores differed across language groups. Steve Wise provides a summary of RTE research and identifies six major insights that the research over the last two decades has achieved.

These four papers seem to cohere elegantly with the six insights Wise has identified.

- **Insight 1**: Effort on ILSAs is not uniform across jurisdictions. OECD nations that encourage student autonomy and tolerate adolescent resistance should not be surprised that their PISA ranks are not as high as jurisdictions that impress the importance of test-taking for national pride and reputation.
- **Insights 2-4**: Problems of method for investigating test effort and engagement are further explicated in the results reported here. Computer-based tests allow identification of multiple effortful behaviours that are not evident in just the amount of time taken to respond. Effort can be seen even when answers are wrong. Use of actions alongside time can create a better way of estimating effortful engagement even when performance is poor; after all performance does not depend solely on effort.
- **Insight 5:** The multi-jurisdictional, multi-lingual comparison revealed that removing RTE scores did not change rank orders, but gaps did get smaller as more potential RTE performances were removed. This suggests strongly to those concerned for jurisdictional reputation that lack of test-taker effort likely explains some of the difference in scores across jurisdictions.
- **Insight 6:** Unsurprisingly, the studies in this special issue did not attempt to monitor or change student effort in an operational PISA test. Wise provides examples of techniques that lend themselves to such monitoring in computer-based testing. The ability to implement such studies will require active collaboration from PISA officials.

Unfortunately, to date PISA seems to have largely resisted any of the suggestions made by academic researchers as to how its score reporting or evaluation of test-taking effort is implemented or how test-taking engagement is achieved. Nonetheless, this special issue provides interesting insights as to how effort and engagement can be operationalised in further research studies. Integration of results across test domains, jurisdictions, item formats, and languages remains to be done. This special issue continues to reveal that effort and engagement do not necessarily change total scores or ranks, but they do demonstrate that PISA scores within jurisdictions are, at least in part, subject to student effort. As such, perhaps educational policy makers should put much less stock in the international rank order comparisons that PISA promulgates. There are better ways to judge the relative merits and characteristics of curricula or systems than depend on international tests in which students might make little effort, in part because they gain so little benefit from participation.

References

Brown, G. T. L. (2008). Students' conceptions of assessment inventory (SCoA version VI) [Measurement instrument]. Auckland, NZ: University of Auckland.

Brown, G. T. L. (2021). Test or invest? NZ's sliding international student assessment rankings are all about choices. *The Conversation*. https://theconversation.com/test-or-invest-nzs-sliding-international-student-assessment-rankings-are-all-about-choices-154729

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92-105. https://doi.org/10.1016/j.compedu.2015.10.018

Zhao, Y. (2014). *Who's afraid of the big bad dragon: Why China has the best (and worst) education system in the world*. San Francisco, CA: Jossey-Bass.

Gavin T. L. Brown
The University of Auckland
gt.brown@auckland.ac.nz
Kane Meissel
The University of Auckland
k.meissel@auckland.ac.nz