ADeLe: Why It Matters and Why the AI Community Should Adopt It

ADeLe Team
https://kinds-of-intelligence-cfi.github.io/ADELE/

Limitations of Benchmark Aggregate Scores for AI Evaluation

Benchmarking with aggregate scores (e.g., accuracy) based on an arbitrary distribution of task items is a useful evaluation paradigm for assessing task-specific performance. This paradigm has been the de facto evaluation approach since the early days of AI, when specialized models were ubiquitous.

It has, however, two main limitations. First, it assumes the test conditions match the deployment conditions, which no longer holds with general-purpose AI (e.g. LLMs), which is expected to be deployed across a diverse and ever-expanding range of tasks in real-world scenarios. Second, it provides limited explanatory and predictive power: the ability to extrapolate benchmark results to new tasks, explaining and anticipating model successes or failures on a case-by-case basis, including unseen real-world challenges beyond what model creators could foresee during the evaluation phase.

Populational Alternatives for Identifying Capabilities

To partially mitigate these two limitations, some efforts resorted to an alternative approach that has been commonly used in psychology and cognitive science for evaluating natural intelligence: Discovering the "structure" of subject (e.g., LLM, human) capabilities, characterized by a small set of underlying capability factors (e.g., language modeling, reasoning, comprehension) that extrapolate to new tasks, using techniques such as 'factor analysis' (Kline, 2014).

Nonetheless, this direct adoption has caveats. Factor analysis has indeed been a successful tool for the extraction of latent capabilities of natural intelligence (e.g., the Cattell–Horn–Carroll framework, the dominant psychometric framework on human intelligence (Carroll, 1993; Schneider and McGrew, 2018)), but the *interpretive validity* of factor-analytic results depends on two preconditions that are much harder to satisfy in AI evaluation: (1) population stability and (2) test diversity.

The first precondition easily holds when evaluating humans because the underlying capability structure of human population is highly stable over time (Carroll, 1993; Wilson et al., 2023), but this is untrue for the rapidly evolving field of AI (Burnell et al., 2023; Ilic, 2023), where we continuously create new systems with increasing capabilities. The reason is simple: if a capability is shared by most models of a generation (i.e., a specific population), it will not explain the variance in that population and that capability will not be extracted as a factor. Similarly, if a capability is lacked by most models of a generation, it will not explain the variance either and will not be captured by methods like factor analysis. The second precondition on test diversity neither comes for free: It requires researchers to meticulously ensure representative coverage and diversity of tests, so as to encompass most (if not all) relevant aspects of human intelligence; this precondition cost decades of active research efforts in psychology (Schneider and McGrew, 2018), and the diversity of aspects and tests that are required for AI may be different from those of humans, in the same way we use different tests for different groups (e.g., children vs adults).

The lack of compliance to the two aforementioned preconditions can be exemplified by the conflicting results in the literature on discovering capability structures of LLMs: Burnell et al. (2023) discovered 3 latent factors (reasoning, comprehension and language modeling), while Ilic (2023) identified one general factor only. In a similar spirit, EpochAI (Burnham, 2025) conducted a principal component analysis (a similar algorithm to factor analysis) and uncovered two latent capabilities in LLMs. Such disagreements primarily stem from (1) the changing populations of AI systems being evaluated over time and (2) relying

on moderate sets of AI benchmarks that lack both coverage and diversity to represent the entirety of cognitive task space. For example, Burnham (2025) used mostly reasoning and math benchmarks.

After all, it is inconsistent to accept that general-purpose models like LLMs only possess 1–3 underlying capabilities and at the same time they present a 'jaggedness' in their capability structure. More importantly, it is inevident how these evaluation results may extrapolate to new benchmarks and real-world deployment conditions; they do not have strong explanatory and predictive power. Better alternatives than factor analysis (and similar approaches like PCA) are needed.

A Promising Alternative: Pre-defined Capability Structures

To this end, a recent *Nature* paper from our team (Zhou et al., 2026) offers an alternative solution: Instead of discovering all latent capability factors a posteriori, we define a moderate set of foundational capabilities *a priori*. This approach predicates that all cognitive tasks, despite their sheer and ever-expanding amount, share 'common grounds'. They may be compressed into a moderate set of foundational capabilities, which may universally manifest (to a greater or lesser extent) across many different kinds of intelligent entities, including humans, animals and machines (Hernández-Orallo, 2017).

To illustrate the promise of this new paradigm, as a first instantiation, Zhou et al. (2026) proposed ADeLe (AI evaluation with Demand Levels) v.1.0, which decomposes the breadth of "intelligence" space into 18 conceptually distinct capabilities that are applicable to LLMs (e.g., reasoning, meta-cognition, comprehension, attention, knowledge of different sciences). Each capability is defined via a multi-experts-derived rubric (grounded in established cognitive and psychometric theories), useful for scalably annotating the extent to which that capability is required (i.e. demanded) to solve any individual task instance.

This unlocks multiple possibilities. First, by combining annotations from all 18 dimensions, we can profile the capability demands of benchmark items (i.e., *what capabilities are required to solve different items inside this benchmark?*), and help diagnose construct validity of benchmarks (i.e., *does this benchmark measure those capabilities that the designers intended to measure?*) at scale for the first time. See Figure 1 for 14 examples of benchmarks where we profiled what capabilities they measure and diagnose construct validity issues.

Second, building on this pre-defined capability structure, it becomes possible to paint a (reasonably holistic) capability profile that summarizes the LLM's intelligence levels, while striking a good balance between coverage, diversity and sample efficiency when selecting benchmark items for evaluation. See Figure 2 for an example where we profile the 18 capabilities of 15 LLMs. It's recommended to check out detailed interpretations of this plot via the original paper for interested reader. This short-to-read, accessible blogpost will focus on the high-level ideas and findings.

Further, we unlock high accuracy on predicting LLM performance when facing completely new tasks and benchmarks, by matching a given LLM's capability profile with the demand profile of a given benchmark or of a specific problem inside that benchmark (demonstrated in Figure 3 in the ADeLe paper).

Last but not least, ADeLe can reconcile conflicting claims about whether AI can or cannot reason, by both controlling confounding capability demands and decomposing reasoning into a multi-dimensional, multi-level space. Overall, the core conclusion is that AI *can* reason, but only up to a point (similar to the way distinct humans or animals manifest varying levels of reasoning ability), and ADeLe can pinpoint exactly where that point is for any given model. For detailed empirical results, we referred readers to consult the full paper.

The Paradigm is Changing, but Not Sufficiently

The first version of ADeLe was published as a preprint in early[1] 2025. We showed how Volume, representing the time a task could take by humans, was a very predictive dimension. A few weeks later, METR published a very similar idea (Kwa et al., 2025), where instead of using a rubric to determine the volume in time that a task could take, they actually measured it in specific tasks where this was feasible, such as software engineering, following the tradition of measuring the length of software projects (successes and failures, Brooks (1974)). This proposal (which resembles our Volume dimension) became arguably the most influential paper in AI evaluation in 2025, with the community using the horizon of

[1] https://arxiv.org/abs/2503.06378.

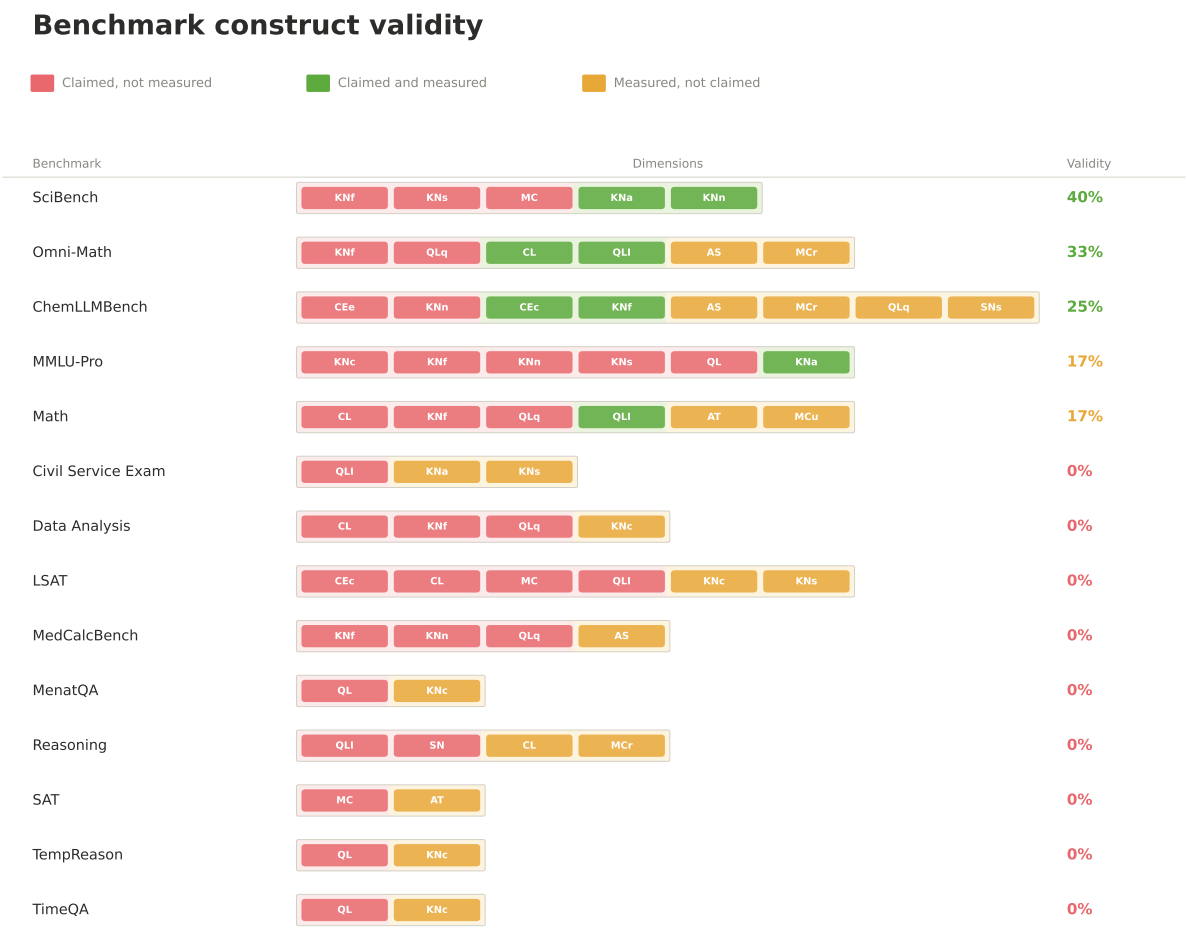

Figure 1: Construct validity analysis of 14 benchmarks.

tasks as an indicator of progress, mostly because it's a very simple (or simplistic[2]) indicator to understand. However, ADeLe has shown that one single dimension gives limited explanatory and predictive power, and this can be substantially improved by considering many other general capability dimensions introduced by ADeLe.

Similarly, item response theory is used more often now. For example, in late 2025, Epoch AI derived a so-called ECI (Epoch Capability Index) metric by applying standard IRT to a collection of models and benchmarks[3], calling this "the Rosetta stone of evaluation" Ho et al. (2025). Nonetheless, unlike ADeLe, it is still populational, inheriting many of the problems of circularity of IRT, the changing scales as the population changes, and the dependency on the benchmark range of difficulty, as mentioned above. While IRT puts the ability of all models in the same scale, this scale is unstable and depends on the other models and benchmarks, which are changing every few months. This limitation is also recently discussed in Anthropic's system card for Mythos[4]. Beyond being populational, ECI is also unidimensional, and thus again gives limited explanatory and predictive power than ADeLe's multidimensional approach.

Some other more recent papers have introduced (or revisited) taxonomies, recognizing that a more complete characterization of AI requires a multidimensional approach, borrowing the term 'capability profile'. Hendryck's 'Definition of AGI" uses Cattell-Horn-Carroll taxonomy and calculates a percentage from these categories Hendrycks et al. (2025). The issues about this 'definition' have been discussed

[2] https://aievaluation.substack.com/p/2025-march-ai-evaluation-digest.
[3] https://epoch.ai/eci/.
[4] https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf

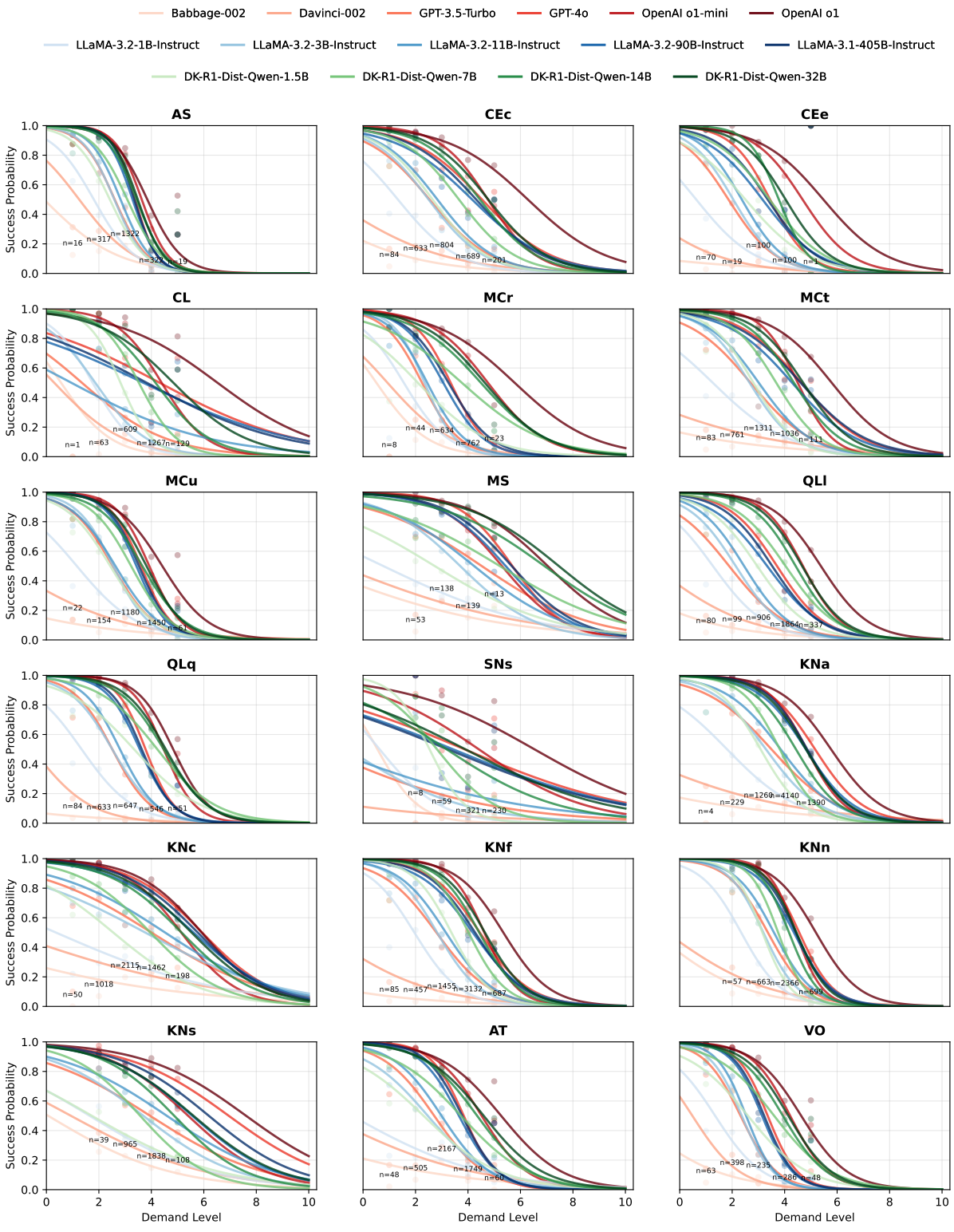

Figure 2: Construct validity analysis of 14 benchmarks. Subject characteristic curves of 15 LLMs for the 18 capability demands. The x-axis shows the demand levels for a given capability dimension and the y-axis the average performance (probability of success) for each level.

elsewhere[5] and compared with ADeLe[6]. A similar, better grounded proposal, is Google DeepMind's

[5] https://aievaluation.substack.com/p/is-the-definition-of-agi-a-percentage.
[6] https://www.gleech.org/ai2025.

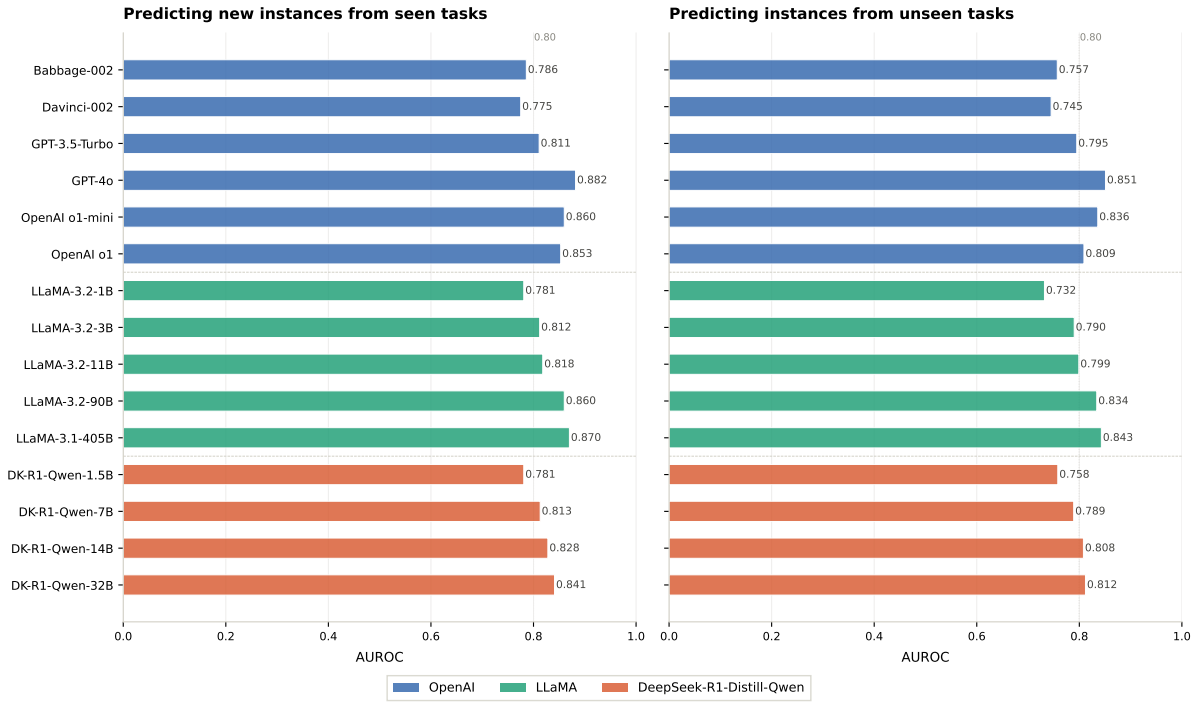

Figure 3: Predictive power unlocked by ADeLe v.1.0 when predicting the performance of 15 LLM subjects from three model families.

proposal[7] released in March 2026, introducing a different theoretical taxonomy. Overall, unlike ADeLe, these approaches do not introduce 'general scales', and thus, again, cannot achieve high explanatory and predictive power.

Perhaps the most notable alternative to ADeLe is the set of OECD scales[8] published in June 2025. The taxonomy for these indicators is also different from the ones we use in the ADeLe paper. What we found is that the initial choice is not that relevant provided you build predictive models to understand the explanatory and predictive power of the taxonomy for each particular model, and can select what dimensions are most relevant for that model and a collection of tasks. The OECD framework does not create the capabilities profiles based on actual experimental results. Instead, the levels are assigned to the models by experts. We think that the intermediate step of ADeLe annotating tasks and then using real results from models to build explanatory and predictive models is much more robust to validate or select from any taxonomy, but we understand the OECD approach is simpler and can give a complementary account of AI capabilities, especially for policy makers.

Looking Ahead

ADeLe is designed to evolve alongside advances in AI, not only for LLMs but can also be extended to multimodal and embodied AI systems. It also has the potential to serve as a standardized framework for AI research, policymaking, and security auditing. More broadly, it advances a more systematic approach to AI evaluation—one explains system behavior and predicts performance.

That said, ADeLe v.1.0 is not without limitations. It was originally developed and pre-printed in March 2025. Since then, AI progress has been notably fast, including the release of much more powerful LLMs and agents. We make a call for the community to join us in the definition of these newer versions of ADeLe, using the best science possible for extracting capability (and propensity) profiles, rather than reinventing the wheel or applying techniques that have many issues.

[7] https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/measuring-progress-toward-agi/measuring-progress-toward-agi-a-cognitive-framework.pdf.

[8] https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/11/oecd-ai-capability-indicators-technical-report_d3762d1a/9cdb3dd1-en.pdf.

References

Frederick P Brooks. The mythical man-month. *Datamation*, 20(12):44–52, 1974.

Ryan Burnell, Han Hao, Andrew RA Conway, and Jose Hernandez Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023.

Greg Burnham. Benchmark scores = general capability + claudiness, 2025. URL https://epoch.ai/gradient-updates/benchmark-scores-general-capability-claudiness. Accessed: 2026-03-28.

John Bissell Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge university press, 1993.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, et al. A definition of agi. *arXiv preprint arXiv:2510.18212*, 2025.

José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

Anson Ho, Jean-Stanislas Denain, David Atanasov, Samuel Albanie, and Rohin Shah. A rosetta stone for ai benchmarks. *arXiv preprint arXiv:2512.00193*, 2025.

David Ilic. Unveiling the general intelligence factor in language models: A psychometric approach. *arXiv preprint arXiv:2310.11616*, 2023.

Paul Kline. *An easy guide to factor analysis*. Routledge, 2014.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring ai ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 352, 2025.

W Joel Schneider and Kevin S McGrew. The cattell-horn-carroll theory of cognitive abilities. *Contemporary intellectual assessment: Theories, tests, and issues*, 733:163, 2018.

Christopher J Wilson, Stephen C Bowden, Linda K Byrne, Louis-Charles Vannier, Ana Hernandez, and Lawrence G Weiss. Cross-national generalizability of wisc-v and chc broad ability constructs across france, spain, and the us. *Journal of Intelligence*, 11(8):159, 2023.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang-Chen, Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, and José Hernández-Orallo. General scales unlock ai evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026.