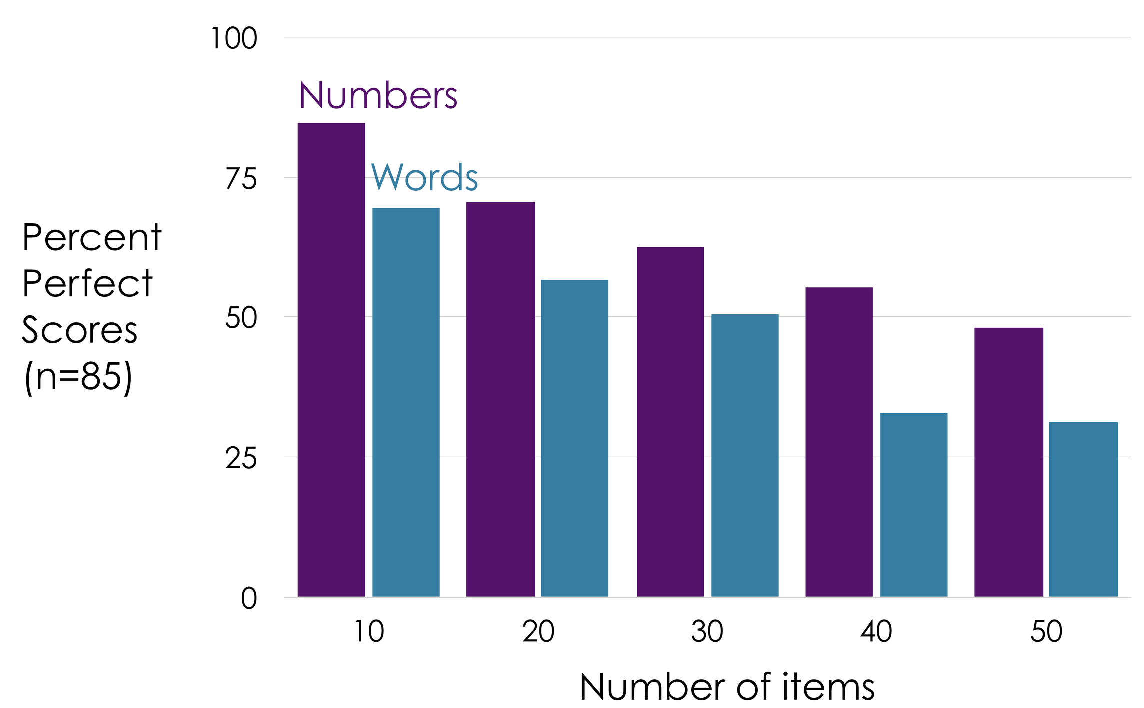
Wise crowds and complex tasks, they're not just for point estimates anymore.

James Heyman & Sandra Rathod
University of St. Thomas

Rankings are great, ranking's a problem.

Ranked preferences contain an extraordinary amount of information (Shannon). Unfortunately, people have trouble ranking more than 10 items. *Declarative* knowledge aside (e.g. Is Iowa bigger than Kentucky?), we have significant *procedural* shortcomings. In terms of algorithmic complexity, ranking is $O(n^2)$; it's hard, we tire, err, and capitulate (Krosnick & Alwin).



The Wisdom of Crowds (WoC) describes how people's aggregated judgements can be more accurate than a crowd's most knowledgeable member. A given crowd's wisdom is a function of their size, individual accuracies and, *most importantly, the heterogeneity of their opinions.*

However, it's WoC works for point estimates but does it work for ranking large sets?

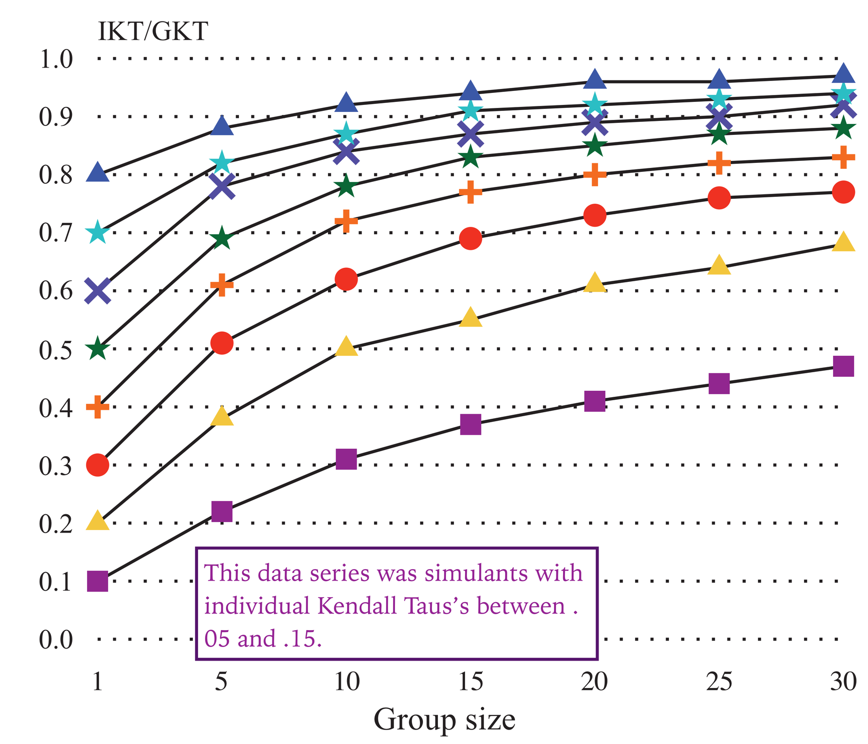
✔ WoC works on point-estimate (PE) problems in countless domains (Surowiecki); all that's needed is a well-structured question, a coherent variety of opinions, and an aggregation method.

✔ PE aggregation is with mean or median — both are understood and well-behaved (Galton). Aggregating rankings is potentially problematic (Arrow); but, as a matter of practice, several methods work (Lee, Steyvers, & Miller).

? PE WoC leverages its 1:1 mapping between judgement diversity and error magnitude; with ranking, this is many: 1. How do the two types of a crowd's variance affects its wisdom?

✘ People's declarative knowledge should combine fine but long-list ranking introduces a lot of non-structured noise that is unrelated to the correct answer. This is akin to people guessing at random.

Simulants ranked 30 items…



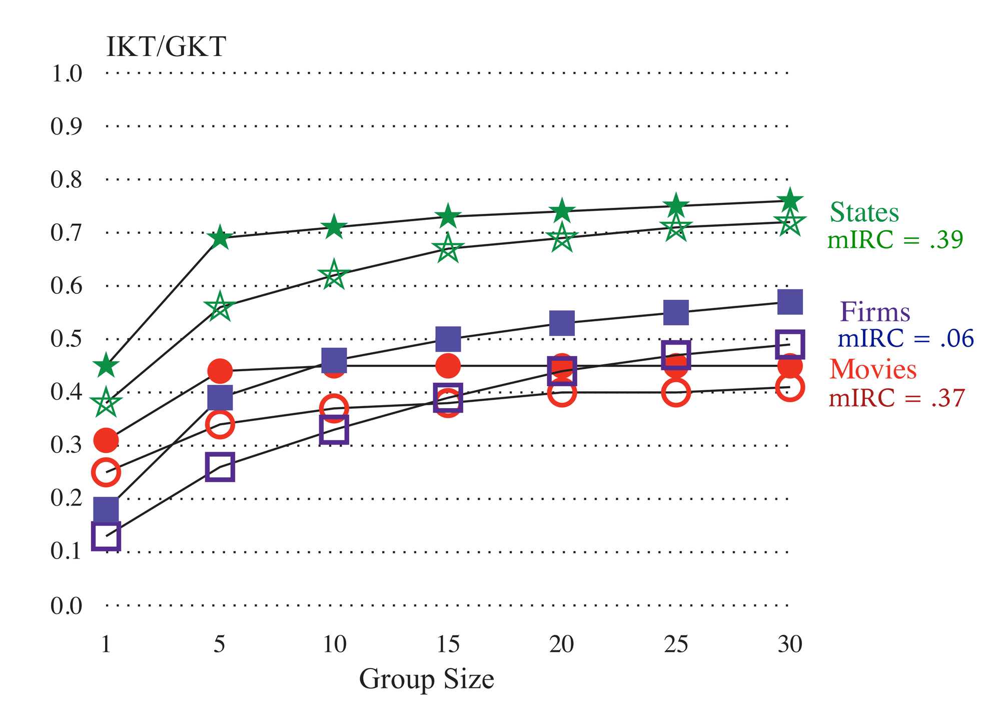
WoC works when applied to people ranking 10 items, however human performance decays rapidly on larger sets so we first tested 30-item ranking with simulants that we then formed into crowds. Unsurprisingly, smarter crowd members made for wiser crowds (Budescu & Chen; Mannes, Soll, & Larrick). More surprising was that for all levels of individual ability, larger crowds helped but had a decreasing marginal effect.

Crowd heterogeneity usually means combining a range of individual abilities (for ranking, IKTs). This Type 1 heterogeneity doesn't predict ranking WoC (presumably)because of the many: 1 judgement:error relationship.

In contrast, using the mean Inter-Ranker Correlation (mIRC) This(Type 2 heterogeneity captures the mix of each crowd's underlying knowledge.

	Model 1		Model 2		Model 3		Model 4	
Variable	B	***B***	B	***B***	B	***B***	B	***B***
IKT	.66	.51	1.30	1.00	1.31	1.01	1.31	1.01
N	.007	.70	.007	.70				
Log(N)					.12	.74	.12	.74
Var(IKT)	.12	.02*						
IRC			-.59	-.53	-.61	-.54	-.60	-.53
3-way							-.02	.-.003$^{.66}$
Adj R^2	.75		.78		.83		.83	

…as did participants.



Students ranked 30 state populations (n=239), Fortune 500 firms' annual revenues (n=238), and movie box office revenues (n=117).

The biggest accuracy gains came from the initial formation of small crowds particularly with States and Movies: domains where people had relatively high levels of knowledge in common.

Paradoxically, it was only in Firms, the domain with low IKTs and low mIRCs, that the benefit of size accrued to larger crowds.

In each domain, smarter people made for wiser crowds although once past an initial boost, larger crowds only mattered when there was a high level of Type 2 heterogeneity. But, Model 4 carried over -- there were no significant interaction effects.

	States			Movies			Firms		
Variable	B	*β*	Effect Size	B	*β*	Effect Size	B	*β*	Effect Size
IKT	1.99	1.42	0.73	1.60	1.00	0.76	2.88	0.84	0.77
Log(N)	0.074	0.55	0.67	0.023	0.22	0.18	0.13	0.62	0.73
IRC	-0.89	-0.87	0.51	-0.45	-0.54	0.43	-1.30	-0.42	0.46
Adj R^2		0.85			0.78			0.86	

WoC improves large-set ranking accuracy.

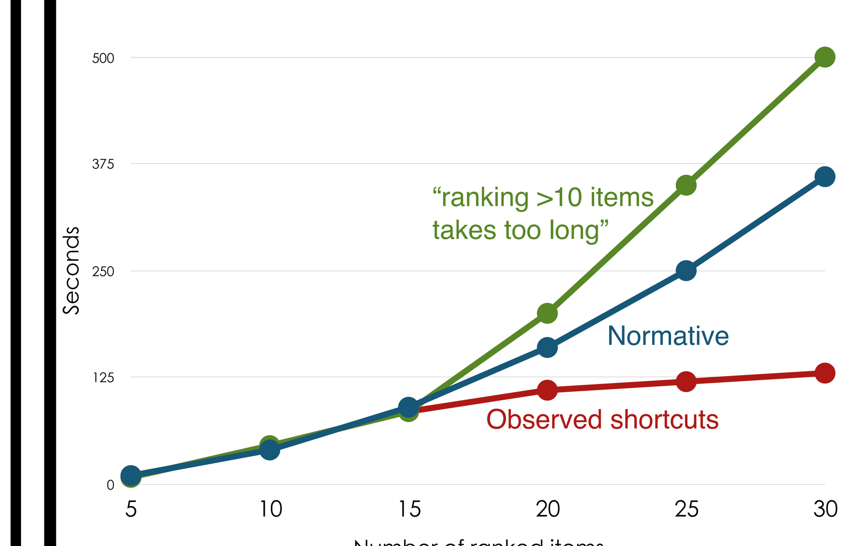
✔ Increases in ranking accuracy are consistent with with those found in point estimate tasks. Applying WoC to ranking corrects for both declarative and procedural deficiencies.

✔ Even small crowds become wiser than their members. Larger crowds can be wiser but size is generally neither necessary nor sufficient.

✔ The effect of Type 2 heterogeneity is consistent with research showing that informational rather than social diversity increases work group performance.

? WoC ranking relies on Type 2 heterogeneity — the differences in people's underlying knowledge rather than simply a dispersal of accuracies. This requires a different analytical approach than point estimates.

A serendipitous finding is that people are better rankers than usually assumed.



Results indicated that people's natural ranking breakpoint is around 15 items rather than <10. Above that point, regardless of issues involving declarative heuristics, they start using procedural shortcuts.

Alwin, D.F., & Krosnick, J.A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49*, 535-552.
Arrow, K.J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy, 58(4)*, 328-346.
Budescu, D.V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science, 61*(2), 267-280.
Davis-Stober, C.P., Budescu, D.V., & Dana J. (2015). The composition of optimally wise crowds. *Decision Analysis, 12*(3), 130-143.
Galton, F. (1907). One vote, one value. *Nature, 75*(1948), 414.
Lee, M.D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PLoS One, 9*(5), 1-9.
Mannes, A.E., Soll, J.B., & Larrick, R.P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276-299.
Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379-423, 623-656, July, October, 1948.
Surowiecki, J. (2005). *The Wisdom of Crowds.* New York: Anchor Books.