Common Retrospective Economic Perception Items Show Good Test-Retest Reliability*

Jack Bailey†

This version March 11, 2022.

Word count: 2,307

Abstract

Retrospective economic perception items dominate economic voting research. Though they are well-used, their measurement properties are not well-known In this short note, I assess the items' test-retest reliability for the first time. I make three contributions. First, I show that the items have good test-retest reliability. Second, I show that personal items are more reliable than national ones. Third, I show that the items' reliability likely does not affect model parameter estimates. Thus, though these items have their problems, reliability is likely not one of them.

*Thank you to Jon Mellon whose advice improved this manuscript.

†Research Associate, Department of Politics, The University of Manchester, UK. If you have any comments or questions, feel free to contact me either by email (jack.bailey@manchester.ac.uk) or on Twitter (@PoliSciJack).

Introduction

One set of items dominates economic voting research. They ask how the national economy and respondents' own personal finances and changed over the past 12 months. When it comes to studying the economic vote, these items are invaluable. Most individual-level data are cross-sectional. They come from only a single country at a single point in time. As such, any macro-economic variation remains constant. The economic perceptions that respondents report, however, do not. We can, therefore, use them to test how respondents' economic views affect their likely vote.

These items are now well-used. But they are not well-understood. What we do know is that they suffer from a host of problems. For instance, they exhibit systematic partisan bias (see Bailey 2019; De Vries, Hobolt, and Tilley 2018; Evans and Andersen 2006). Here, incumbent supporters tend to be more positive, and opposition supporters more negative, than similar non-partisans. Likewise, respondents also report different perceptions in different contexts (Bailey 2021) and change the perceptions they report depending on preceding items (Wilcox and Wlezien 1993; Sears and Lau 1983). Though we now have a good idea about these issues, we still do not know much about the items' measurement properties. In particular, how prone they are to measurement error.

In this short note, I assess the test-retest reliability of common retrospective economic perception items for the first time. Test-retest reliability is an essential metric when it comes to survey research. Imagine that we repeat an item twice, only a short time apart. A reliable item would show almost exactly the same set of responses each time. An unreliable item would not. Either way, we learn something useful: the item's susceptibility to measurement error. We can then use this information to improve the measurement properties of our surveys and, thus, any subsequent inferences.

I make three substantive contributions. First, I show that these items have a good level of reliability. When asked to report their perceptions twice in a single survey, most respondents give the same answer each time. Second, I show that personal items are more reliable than national

ones. This makes sense, as respondents are likely more familiar with their own circumstances than with the state of the national economy. Third, I show that measurement error does little to affect estimates from conventional economic voting models. As such, I conclude that common retrospective economic perception items have good test-retest reliability.

Data

My data come from wave 22 of the British Election Study Internet Panel (Fieldhouse et al. 2021). Fieldwork occurred between 26 November 2021 and 15 December 2021. Further, the survey took around 15 minutes to complete on average.

Most waves of the British Election Study Internet Panel measure respondents' economic perceptions only once. This is, of course, to be expected. But wave 22 instead measured them *twice*. The first time was as usual, with all 28,135 respondents reporting their perceptions. The second time was around 5 minutes later, when a subset of 6,948 respondents reported them for a second time. It is this subset of respondents that I take as my sample.

Assessing Test-Retest Reliability

Assessing test-retest reliability is straight-forward. First, one fields an item to a sample of respondents. After fielding any interim items, one then fields the item of interest in exactly the same way a second time. This provides two responses to the item. Armed with each response, one can then estimate their correlation to get the item's test-retest reliability (Yu 2005).

This process requires two assumptions. First, one must assume that respondents' latent perceptions remain stable throughout the survey. As it took my respondents only around 15 minutes to complete, this seems reasonable. Second, one must assume that each response instance measures respondents' latent perceptions *with error*. Again, as measurement error is an unavoidable factor in survey research, this seems reasonable too.

Table 1 shows each item's test-retest reliability across four different metrics. Note that both items are nominal as they include a "Don't know" option. As such, Cramér's V (Cramér 1946)

Table 1: Measures of test-retest reliability across two responses to each retrospective item embedded in wave 22 of the British Election Study Internet Panel

Item	Repeated	Cramér	Pearson	Spearman
National Economic Perceptions	72.3%	0.60	0.73	0.75
Personal Economic Perceptions	87.3%	0.79	0.83	0.88

is the most appropriate metric to use[1]. Cramér's V provides a measure of association between nominal variables. Still, the three other metrics – the percentage of repeated responses and the items' Pearson/Spearman correlations – produce similar results[2].

Both items appear reliable. In general, respondents tend to give the same answer each time. Note that the national item is less reliable than the personal one across all four metrics. Consider Cramér's V. Here a score of 0 implies no association and a score of 1 implies total association. The personal item has a score of 0.79. But the national item scores only 0.60. This difference in reliability is perhaps unsurprising. After all, we would expect most people to have a better understanding of their own finances than of the state of the national economy.

We can learn more about the items by considering them in greater detail. Figure 1 shows the joint response distribution across both instances for each item. Here, columns reflect the first response and rows reflect the second. Likewise, darker colours imply a greater level of reliability from one instance to the next.

The figure makes three facts most clear. First, responses change only a little between instances, if at all. As we would expect, cells on the diagonal include the highest scores. But the cells with the next highest scores are almost always those next to the diagonal. For instance, 25.5% of respondents who said that the economy "got a lot worse" went on to say that it "got a little worse." Likewise, 20.4% of respondents who said that their own personal finances "got a lot worse" did the same.

Second, responses to the national item appear much less stable at the "better" end of the scale than do responses to the personal item. Only 50.0% of respondents who said that the economy

[1] To help the reader calibrate their understanding of this metric, I provide wave-on-wave estimates of Cramér's V across the entire British Election Study Internet Panel in figure A1

[2] The latter two metrics apply to continuous and ranked data, respectively. As such, I remove any "Don't know" responses in these cases.

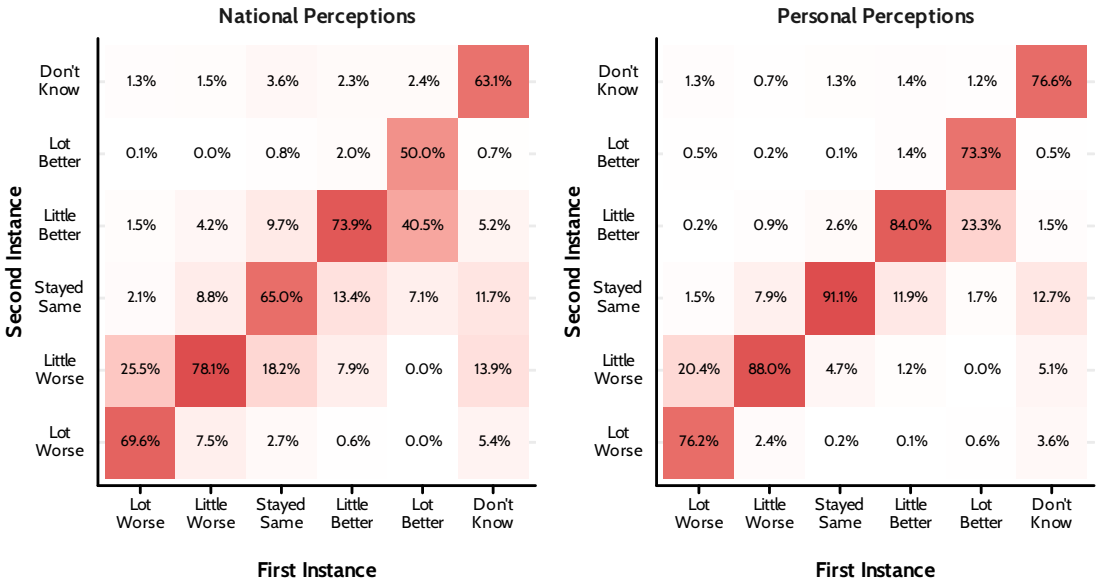

Figure 1: Crosstabs comparing first (columns) and second (rows) responses to both retrospective national and personal economic perceptions items in wave 22 of the British Election Study Internet Panel. Figures show column percentages.

"got a lot better" did so a second time. This compares to 73.3% of respondents who said that their own personal finances "got a lot better." Again, this makes sense if respondents have a better idea of their own finances than of the national economy. But it also makes sense given that the British economy was still reeling from the economic onslaught of the covid-19 pandemic. As a result, it might not be surprising that respondents did not repeat their strong "got a lot better" responses.

Third, respondents often appear not to be telling the whole truth when they say "don't know." Instead, "don't know" often appears to mean "can't be bothered to answer." This is true for both items. For example, 36.9% of respondents who gave a "don't know" response to the national item and 23.4% who gave a "don't know" response to the personal item then went on to provide an informative answer on their second attempt.

Consequences of Measurement Error

My results show that common retrospective economic perception items are broadly reliable. That said, they do include some measurement error. Further, this error is not constant across items or response options. Thus, these differences also yield different economic voting estimates.

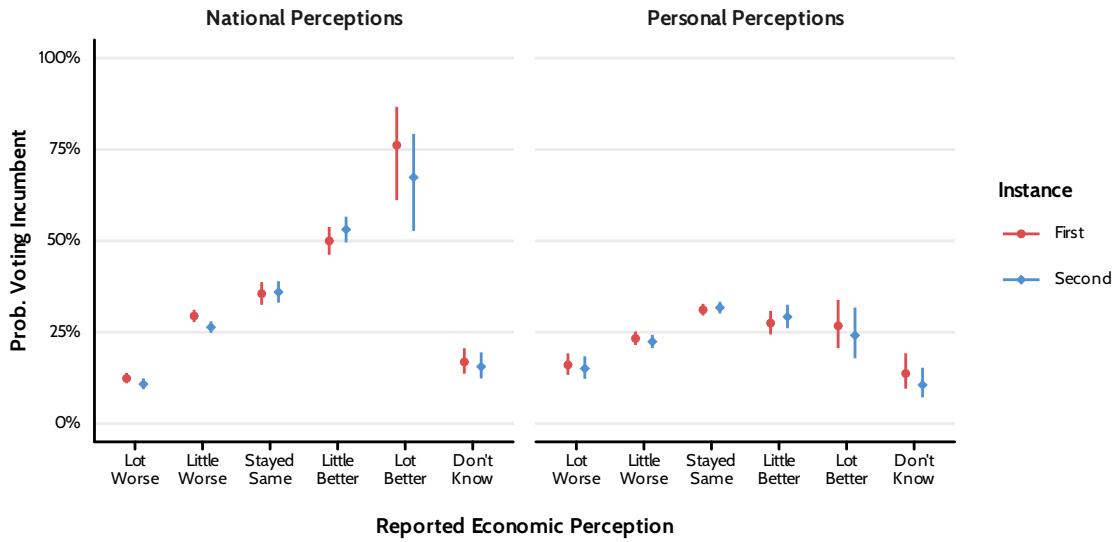

Figure 2: Despite some measurement error, the bivariate relationship between reported economic perceptions and incumbent support remains broadly stable across both response instances

Individual-level economic voting models tend to be quite simple. Most often, they treat incumbent voting as a function of socio-demographics and reported economic perceptions. I fit four such models below. Each corresponds to each instance of my two items. Economic voting research almost always treats these scales as continuous. I do not, for two reasons. First, that doing so requires less strict assumptions of linearity. Second, that doing so reflects the fact that each response option should include different amounts of error.

Figure 2 shows predictions from each model. Other than some minor differences, models fit to each item produce similar results. Though the estimates that they produce differ, they almost always show considerable overlap[3]. This suggests an interesting possibility. Duch and Stevenson (2008) note that economic voting estimates show some variability. Measurement error may, thus, be a leading cause of this variation.

As an aside, it is worth noting that the personal item has a non-monotonic effect on incumbent voting. Most applied economic voting research treats this item as continuous. But, clearly, its effect is non-linear. As such, past research may not tell us the whole picture. At the negative end of the scale, this research will understate the item's influence. Likewise, at the

[3]Note that I do not mean to imply that both estimates are bias free. Indeed, as we know that each item suffers from measurement error the opposite is likely true: measurement error will induce attenuation that biases any effects towards zero. As such, I mean only to imply that both instances produce similar estimates, no matter their flaws.

positive end of the scale it will overestimate its influence. Economic voting scholars should, thus, consider relaxing any strict linearity assumptions.

Conclusion

Common retrospective economic perception items have their issues. Most notably, partisan bias and survey-induced measurement problems. Both pose serious issues for the validity of many economic voting findings. And the results that I present in this short note do nothing to change this fact.

Even so, my results do at least show that these items are reliable. When respondents report their perceptions minutes apart, they differ only little. What's more, these reported perceptions also yield similar parameter estimates and inferences. Thus, despite their issues, these items do produce repeatable results and inferences.

My findings quantity the measurement error present in retrospective economic perceptioin items for the first time. Though this measurement error is not large, it remains a nuisance nonetheless. In particular, it causes attenuation bias that pulls parameter estimates towards zero. Reducing this error, and thereby reducing attenuation bias, is, thus, an important challenge.

One way to make economic voting research more robust would be to use multi-item scales. With only one item, it is not possible to separate signal from noise. But this is possible with more than one item. For instance, one could estimate noise-free latent perceptions using IRT (de Ayala 2009) or factor analysis (Brown 2015). Multi-item scales would present new opportunities too. At present, we treat economic perceptions as unidimensional. But this might not be the case. For example, it might matter not only what voters *believe* about the economy, but how it makes them *feel* too. New items, thus, offer both better measurement properties and a chance to reinvigorate economic vote.

References

Bailey, Jack. 2019. "The Fact Remains: Party ID Moderates How Voters Respond to Economic Change." *Electoral Studies* 61 (102071): 1–13. https://doi.org/10.1016/j.electstud.2019.102071.

———. 2021. "Political Surveys Bias Self-Reported Economic Perceptions." *Public Opinion Quarterly* 85 (4): 987–1008. https://doi.org/10.1093/poq/nfab054.

Brown, Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research*. Second. New York City, NY: Guilford Publications.

Cramér, Harald. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

de Ayala, R J. 2009. *The Theory and Practice of Item Response Theory*. New York, NY: The Guildford Press.

De Vries, Catherine E., Sara B. Hobolt, and James Tilley. 2018. "Facing up to the Facts: What Causes Economic Perceptions?" *Electoral Studies* 51 (February): 115–22. https://doi.org/10.1016/j.electstud.2017.09.006.

Duch, Raymond M., and Randolph T. Stevenson. 2008. *The Economic Vote: How Political and Economic Institutions Condition Election Results*. Cambridge: Cambridge University Press.

Evans, Geoffrey, and Robert Andersen. 2006. "The Political Conditioning of Economic Perceptions." *The Journal of Politics* 68 (1): 194–207. https://doi.org/10.1111/j.1468-2508.2006.00380.x.

Fieldhouse, Edward, Jane Green, Geoffrey Evans, Jonathan Mellon, Christopher Prosser, Roosmarijn A. de Geus, and Jack Bailey. 2021. "British Election Study Internet Panel, 2014-2023."

Sears, David O., and Richard R. Lau. 1983. "Inducing Apparently Self-Interested Political Preferences." *American Journal of Political Science* 27 (2): 223–52. https://doi.org/10.2307/2111016.

Wilcox, Nathaniel, and Christopher Wlezien. 1993. "The Contamination of Responses to

Survey Items: Economic Perceptions and Political Judgments." *Political Analysis* 5: 181–213.

Yu, Chong Ho. 2005. "Test-Retest Reliability." In *Encyclopedia of Social Measurement*, edited by Kimberly Kempf-Leonard, 777–84. New York: Elsevier. https://doi.org/10.1016/B0-12-369398-5/00094-3.

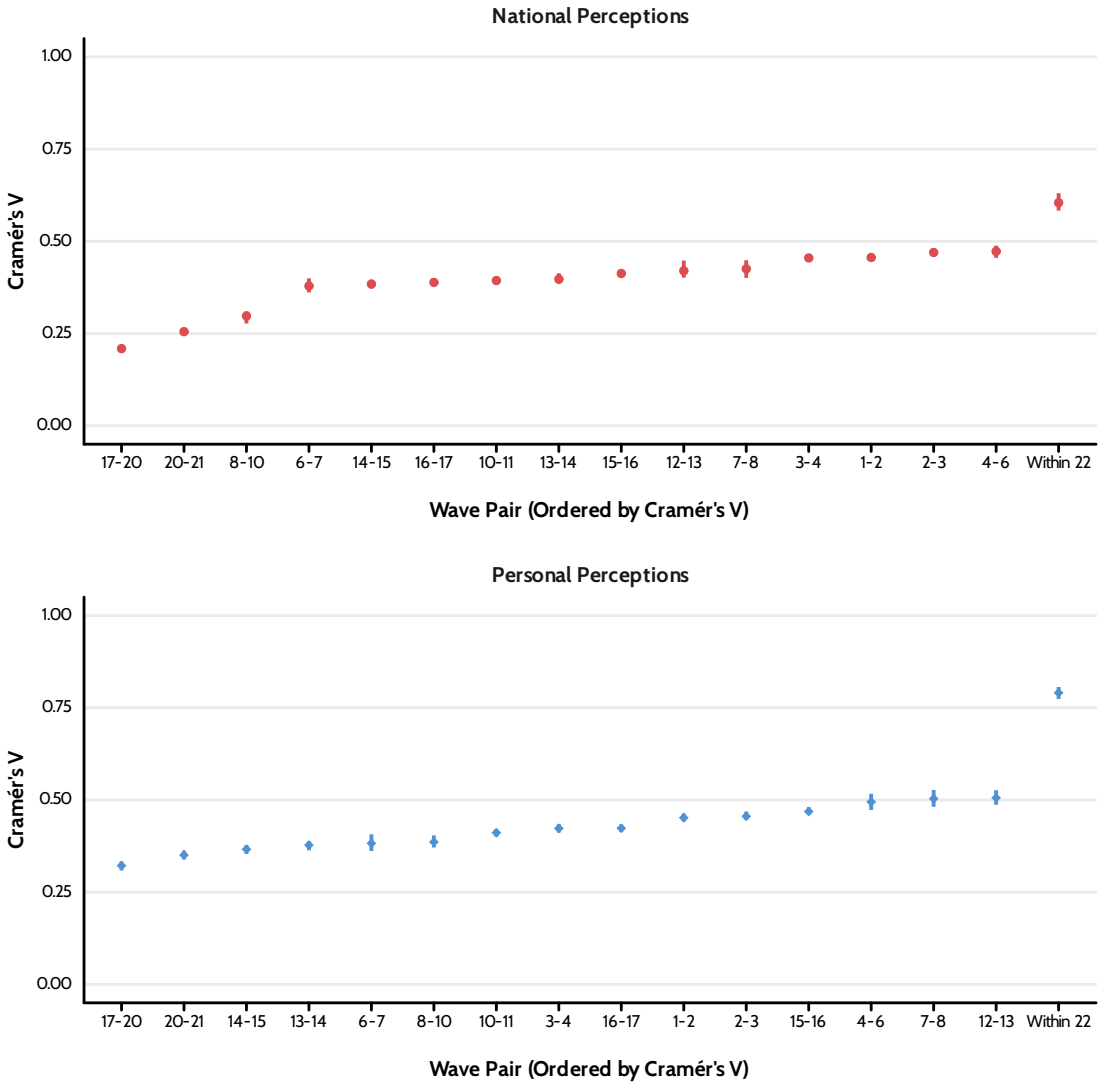

Figure A1: The bivariate relationship between reported economic perceptions and incumbent support remains broadly stable across both response instances and both items