The Reliability of Survey Measures

RESULTS Series

COMPARISON OF HEISE 3-WAVE SIMPLEX WITH TEST-RETEST RESULTS

Paula A. Tufiș

University of Bucharest

Duane F. Alwin

Pennsylvania State University and the University of Michigan

JUNE 2023

Suggested citation: Tufiș, P.A. & Alwin, D.F. (2023). Comparison of Heise 3-wave Simplex with Test-retest Results. *The Reliability of Survey Measures Results Series.*

Here[1] we report reliability estimates for approximately 600 measures in the three GSS panels. We focus on non-redundant, self- and proxy reports only; excluding performance measures, as well the as eliminating interviewer and organization reports. In our Appendix table we present a summary of our findings for each distinct question in the pool of GSS items considered here, averaged over common items in the pool.

Stability of Measures

In addition, we also present the 4-year stability of the underlying trait, quantifying the extent to which there is true change in the underlying trait being measured assessed at the population level. The stability estimate is based on Heise's (1969) formula, specifically $CR(13)^2 / CR(12) * CR(23)$ (see equation 12, page 97). [2] These 4-year stability estimates range from high levels, i.e. 1.0, to relatively lower levels.

[1] This document is an extract from the research paper "A Catch-22—the Test Retest Method of Reliability Estimation," by Paula A. Tufiş, Duane F. Alwin, and Daniel N. Ramírez. Table and figure numbers refer to those in the parent paper.

[2] As depicted in Figure 1, there were a small number of cases where the stability exceeded the theoretical limit of 1.0 (standardized). We eliminated standardized stabilities that exceeded 1.15 (11 cases), and we set those stabilities falling between 1.0 and 1.15 to 1.0.

Table 1. Reliability estimates and differences by stability, averaged over GSS panels, for non-redundant self- and proxy-reports

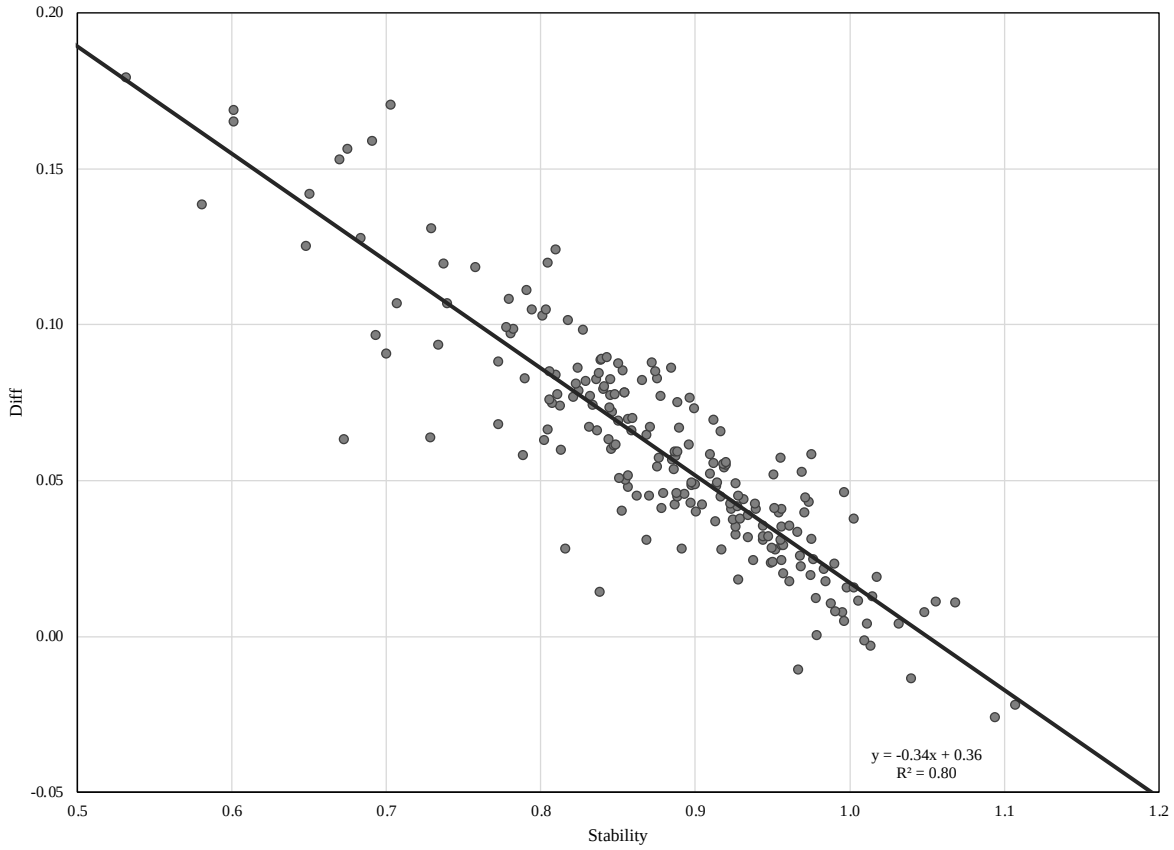
	Number of measures	Stability	TRT	Heise	Diff	t-test	df	p-value
Fixed traits	11	.975	.856	.872	.016	3.258	10	.009
Highly stable traits (stability = .93 - 1.0)	52	.964	.744	.764	.020	8.373	51	.000
Relatively stable traits (stability = .87 - .92)	54	.903	.672	.720	.048	29.120	53	.000
Less stable traits (stability = .82 - .86)	52	.846	.604	.673	.069	26.741	51	.000
Unstable traits (stability < .82)	53	.745	.496	.595	.099	19.353	52	.000

Notes: Fixed traits: cohort, agekdbrn, fund16, granborn, madeg, maeduc, mapres80, padeg, paeduc, papres80, incom16. Stabilities over 1.15 are coded as missing; stabilities between 1 and 1.15 are coded as 1. TRT, Heise, stability and difference estimates are averaged over common items in the pool.

We present a summary of these results in Table 2. In general, as expected the 3-wave Heise estimate is greater than CR(21), the TRT estimate, although there are a substantial number of cases in which the estimates are virtually identical.[3] In this table results are presented for several categories of measures ordered by levels of stability, including a small set of questions that are "fixed" in the sense that they inquire about traits that theoretically cannot change (e.g., birth year), and for quartiles of the 4-year stability estimate. Hout and Hastings (2016) have already demonstrated the high levels of reliability with these fixed questions. As indicated in this summary table, we performed a test of the difference between the TRT and the Heise estimates, using a test of "matched pairs" (see Blalock 1972, pp. 233-235). These results indicate that for "fixed" traits, or for highly stable traits, the differences between the two estimates are small and not statistically significant at the $p < 0.001$ level. As the extent of change in the underlying trait increases, that is, as instability increases, the differences are greater and statistically significant. These patterns are depicted in Figure 7, where we present the resulting scatterplot relating the difference score [i.e. HEISE minus TRT to the level of stability, and the linear regression of the difference on stability ($R^2 = .80$). The results summarized here clearly suggest that the difference between the estimates is in part a function of the stability of the trait being measured. Kiley and Vaisey's (2021) results anticipate the fact that many of the GSS questions reveal high levels of stability.

[3] There was a small number of cases where the test-retest estimate was greater than the Heise estimate, that is, the Heise minus TRT value was negative, although in these instances the differences were very small. With a fair degree of confidence, we attribute these differences to sampling error, and for our present purposes, in Table 2 we set these differences to zero.

Figure 1. Scatterplot of the relationship between the difference score (Diff) and the stability



Content of Measures

In addition to the stability of the trait involved, one of the possible factors that contributes to the disparity between the two approaches is the nature of the *content* being assessed. By content, we refer to whether the variable of interest is a fact or non-fact, and the type of non-fact (i.e. subjective assessments) being measured, specifically non-facts involving beliefs, attitudes, values, self-perceptions, self-evaluations, or expectations. There is a well-established finding in the survey methods literature that the measurement of factual content (objective information that can be verified) can be assessed more accurately than non-facts in survey reports (Alwin 2007; Hout and Hastings, 2016). Thus, we hypothesized that the content being measured may affect the differences between the two reliability estimates.

In order to examine this hypothesis, we present the mean estimates of reliability for self- and proxy-reports, averaged across GSS panels, organized by question content and the approach to reliability estimation. This table permits us to analyze the differences between the TRT and Heise, or Quasi-Markov Simplex Models (QMSM) estimates within categories of content. Question content is operationalized here according to Alwin's (2007, pp. 153-154) differentiation of facts (content that can be verified), vs. non-facts (which are largely subjective states), as well as differences among types of non-factual content: beliefs, attitudes, values, self-perceptions, self-assessments and expectations (see Alwin 2007, pp. 153-154, for a detailed discussion of these).

Table 2. Mean estimates of reliability, by question content and approach to reliability estimation, averaged across GSS panels, for non-redundant self- and proxy-reports

				TRT - Heise Comparisons		
Content	Measures	TRT	Heise	t test	df	p-value
Facts	35	.797	.847	6.617	34	.000
Non-facts	176	.595	.656	22.915	175	.000
Beliefs	67	.568	.634	13.564	66	.000
Values	42	.612	.670	12.337	41	.000
Attitudes	35	.616	.671	10.455	34	.000
Self-Assessments	12	.579	.652	6.406	11	.000
Self-Perceptions	14	.691	.740	8.558	13	.000
Expectations	6	.476	.532	3.167	5	.025
Total	211	.629	.688	23.150	210	.000
Comparisons						
All content						
F-ratio		13.077	13.022			
p-value		.000	.000			
Facts vs. Non-facts						
F-ratio		59.875	60.955			
p-value		.000	.000			
Within Nonfacts						
F-ratio		2.716	2.395			
p-value		.022	.040			

The results in Table 3 provide a formal test of the differences within categories of content, specifically fact vs. non-fact, and within types of non-facts. We employ the "paired samples" t-test procedure used above (see Blalock 1972), which compares the means of two variables for a single group—in this case, the two variables are the test-retest and Heise simplex (QMSM) estimates of reliability—testing whether the average differences in the estimates of the two approaches differ from 0.0. These results consistently reveal systematic differences between the two approaches to reliability estimation, with the Heise simplex estimates averaging at higher levels.

Consistent with prior research, the results in Table 3 also demonstrate that questions involving subjective content have lower reliabilities, a well-established finding in the literature (Alwin 2007; Hout and Hastings, 2016). There are some differences in average reliability across types of non-factual content; there are some demonstrable differences here that coincide with previous results (Alwin 2007, pp. 158-162). These results indicate there are some significant differences (at the $p < 0.05$ level) between content within non-facts. Self-assessments and self-perceptions have the highest levels of reliability, and expectations are measured with least reliability. Both approaches to reliability estimation reveal these same patterns.

Stability vs. Content

We further examine the relationship between stability and reliability estimates using linear regression to summarize our findings. Table 4 presents a series of regression models that summarize our results and parameterize the effects of several predictor variables on the difference between the two estimates (i.e. Heise minus TRT).

Table 3. Regression of GSS reliability estimates on attributes of questions: pooled GSS panels

	Model [1]				
Predictors	1	2	3	4	5
Intercept	.712 ***	.070 ***	.138 ***	.062 ***	.071 ***
TRT (centered)	.892 ***	–	–	–	–
Stability (centered) [2]		-.039 ***	–	–	-.040 ***
Stability quartiles [3]					
2nd quartile			-.060 ***	–	–
3rd quartile			-.089 ***	–	–
4th quartile			-.120 ***	–	–
Content: fact versus non-fact [4]					
Non-facts--beliefs				.022 **	-.003
Non-facts--values				.015 *	.005
Non-facts--attitudes				.010	-.001
Non-facts--self assessments				.026 *	-.006
Non-facts--self perceptions				.002	-.003
Non-facts--expectations				.022	-.016 *
R^2	.877	.785	.636	.054	.790
N of cases	594	594	594	594	594

Key: † p ≤. 10 *p ≤ .05 **p ≤ .01 ***p ≤ .001

[1]Panel fixed effects included (not shown). The first panel is the reference category

[2] Stability is expressed as units of 0.10

[3] 1st (lowest) stability quartile is reference group

[4] Fact category is reference group

Model 1: Regress Heise reliability on TRT-reliability
Model 2: Regress Heise-TRT Difference on Stability (centered)
Model 3: Regress Heise-TRT Difference on Stability as quartiles
Model 4: Regress Heise-TRT Difference on Facts vs. type of non-facts
Model 5: Regress Heise-TRT Difference on Stability (centered) and Content

Note: In Model 1 the regressand is the Heise estimate.
Note: In Models 2-5 the regressand is the Heise-TRT Difference score.
Note: In Model 4 and 5 "facts" is the omitted category

The first model in this table reveals the convergences between the two estimates of reliability. The relationship between the two estimates is high ($R^2 = .877$), but this does not mean they are identical. The remaining models in Table 4 regress the difference (i.e. Heise – TRT) on these factors. As revealed in model 2 of Table 4, the difference is highly predictable from the 4-year stability estimate. This model establishes the linear relationship we previously presented in Figure 7, and the use of quartiles of the stability distribution in model 3 reinforces the finding that the relationship is linear. [4]

In model 4 we regress the difference between the two estimates on facts vs. non-facts, employing a set of dummy variables to represent the types of non-facts (note that the omitted category in this regression is facts). These results indicate that there is a significant difference between facts and non-facts in the difference between the estimates, indicating that the difference is significantly less for facts relative to three categories of non-facts, specifically beliefs, values, and self-assessments. All other types of non-facts are not significantly different from facts. Finally, in model 5 we regress the difference between the estimates on the content (facts vs. non-facts) dummies, while controlling for stability. These results indicate that the content effect is spurious, once stability is controlled, given that facts are mostly highly stable traits. Except for the small negative effect of expectations in the pooled data, there are no substantive differences due to content, once stability of the underlying trait is controlled.

[4] We tested the nonlinear form of this model using loess curves and formally testing the inclusion of a quadratic term. The results showed the optimal fit to the data was linear, indicating there was no need for a quadratic term in the regression in model 2.

References

Alwin, D. F. (2007), *Margins of Error—A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley & Sons, Inc. [Wiley Series in Survey Methodology]

Blalock, H. M., Jr. (1972), *Social Statistics*. New York: McGraw-Hill.

Heise, D. R. (1969), "Separating Reliability and Stability in Test-retest Correlation," *American Sociological Review,* 34, 93-191.

Hout, M., and Hastings, O. P. (2016), "Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006-2014," *Sociological Science* 3, 971-1002.

Kiley, K, and Vaisey, S. (2020), "Measuring Stability and Change in Personal Culture Using Panel Data," *American Sociological Review,* 85(3), 477-506.

Appendix Table 1. Reliability estimates by each GSS item, averaged over panels

Var	TRT	Heise	Diff	Stability	Nr. panels	Var	TRT	Heise	Diff	Stability	Nr. panels
inequal3	.466	.440	-.026	1.094	1	**speduc**	.893	.922	.029	.957	3
librac	.560	.538	-.022	1.107	3	**trust2**	.801	.831	.031	.955	3
degree	.901	.887	-.013	1.040	3	**helpful2**	.705	.736	.031	.944	3
incgap	.478	.468	-.010	.967	1	**socfrend**	.510	.541	.031	.868	3
agekdbrn	.936	.933	-.003	1.013	3	**educ**	.882	.914	.031	.975	3
padeg	.941	.940	-.001	1.009	3	**polviews**	.638	.670	.032	.934	3
spfund	.855	.855	.001	.979	3	**getahead**	.444	.476	.032	.944	3
fehire	.450	.454	.004	1.032	3	**god**	.813	.846	.032	.947	3
maeduc	.873	.877	.004	1.011	3	**cappun**	.854	.886	.033	.926	3
cohort	.990	.995	.005	.996	3	**helpblk**	.593	.627	.034	.966	3
granborn	.960	.968	.008	.995	3	**pray**	.818	.853	.035	.926	3
coneduc	.472	.480	.008	1.048	3	**racdif2**	.643	.679	.035	.956	3
fepol	.688	.696	.008	.991	3	**rellife**	.647	.683	.036	.961	1
paeduc	.920	.931	.011	.988	3	**abnomore**	.835	.871	.036	.944	3
discaff	.397	.408	.011	1.068	3	**grass**	.874	.911	.037	.913	3
letin1	.546	.557	.011	1.056	3	**papres80**	.756	.794	.038	.924	1
divlaw2	.833	.844	.011	1.005	3	**liveblks**	.377	.414	.038	1.002	3
polhitok	.748	.760	.012	.978	3	**spkcom**	.780	.818	.038	.929	3
mapres80	.758	.770	.013	1.015	1	**incom16**	.555	.594	.039	.934	3
discaffm	.350	.365	.014	.838	3	**postlife**	.877	.917	.040	.954	3
colhomo	.750	.766	.016	.998	3	**sexeduc**	.789	.829	.040	.970	2
suicide1	.781	.797	.016	1.003	3	**natpark**	.470	.510	.040	.900	3
abdefect	.843	.860	.018	.961	3	**rincom06**	.751	.792	.041	.852	2
gunlaw	.658	.676	.018	.984	3	**fund16**	.829	.870	.041	.939	3
finrela	.592	.611	.018	.928	3	**conbus**	.488	.529	.041	.923	3
natracey	.633	.652	.019	1.017	3	**colath**	.640	.681	.041	.956	3
childs	.891	.911	.020	.975	3	**closewht**	.458	.499	.041	.878	3
class	.682	.702	.020	.957	3	**discaffw**	.367	.408	.041	.951	3
natcityy	.473	.495	.022	.983	3	**fair2**	.757	.799	.042	.927	3
abany	.830	.852	.023	.969	3	**marblk**	.599	.641	.042	.887	3
madeg	.899	.922	.023	.990	3	**popespks**	.578	.620	.042	.904	3
fund	.852	.876	.024	.949	3	**spanking**	.658	.700	.043	.922	3
nataidy	.641	.665	.024	.950	3	**abpoor**	.838	.881	.043	.939	3
workblks	.341	.365	.025	.937	3	**letdie1**	.780	.823	.043	.897	3
reborn	.899	.924	.025	.956	3	**pornlaw**	.586	.630	.043	.973	3
absingle	.834	.859	.025	.977	3	**abhlth**	.843	.887	.044	.931	3
spdeg	.909	.935	.026	.968	3	**libath**	.599	.643	.045	.971	3
homosex	.876	.904	.028	.952	3	**wrkwayup**	.587	.632	.045	.916	3
racdif1	.690	.718	.028	.917	3	**suicide3**	.773	.818	.045	.888	3
wlthblks	.309	.337	.028	.816	3	**tax**	.635	.680	.045	.862	3
punsin	.602	.631	.028	.891	1	**marwht**	.370	.416	.045	.870	3
fepresch	.540	.569	.029	.949	3	**abrape**	.865	.910	.045	.928	3
premarsx	.783	.812	.029	.956	3	**fefam**	.606	.651	.046	.893	3

Var	TRT	Heise	Diff	Stability	Nr. panels	Var	TRT	Heise	Diff	Stability	Nr. panels
parsol	.611	.657	.046	.888	3	**relactiv**	.639	.705	.066	.836	3
affrmact	.600	.646	.046	.879	3	**conmedic**	.488	.554	.066	.804	3
polattak	.500	.546	.046	.996	3	**nathealy**	.508	.574	.067	.890	3
richwork	.711	.759	.048	.856	3	**obey**	.597	.664	.067	.871	3
partyid2	.861	.910	.048	.913	3	**natcrimy**	.603	.670	.067	.831	3
aged2	.675	.724	.049	.898	3	**jobfind**	.629	.697	.068	.773	3
marhomo	.787	.836	.049	.900	3	**fechld**	.526	.596	.069	.850	3
reliten2	.863	.912	.049	.926	3	**spkmil**	.627	.696	.070	.912	3
xmovie	.807	.857	.049	.897	3	**closeblk**	.592	.662	.070	.856	3
bible	.746	.796	.050	.914	3	**conlabor**	.518	.588	.070	.859	3
marasian	.485	.535	.050	.855	3	**prestg80**	.702	.774	.072	.846	1
conarmy	.563	.614	.051	.851	3	**uswary**	.665	.738	.073	.899	3
marhisp	.494	.546	.052	.857	3	**colcom**	.622	.696	.073	.844	3
life	.602	.654	.052	.950	3	**hrs1**	.512	.587	.074	.812	3
suicide4	.752	.804	.052	.910	3	**kidssol**	.605	.680	.074	.833	3
natdrug	.419	.472	.053	.969	3	**natsci**	.471	.546	.075	.807	3
sprtprsn	.746	.800	.054	.886	3	**racdif4**	.624	.699	.075	.888	3
relpersn	.772	.826	.054	.919	3	**natspacy**	.659	.735	.076	.806	3
raclive	.796	.850	.055	.875	3	**racopen2**	.580	.657	.077	.896	3
prayer	.706	.761	.055	.920	3	**socrel**	.510	.587	.077	.821	3
fear	.697	.752	.055	.918	3	**happy**	.515	.592	.077	.832	3
polescap	.554	.609	.056	.912	3	**courts2**	.784	.861	.077	.877	3
helpnot	.453	.509	.056	.920	3	**income06**	.803	.881	.077	.845	3
conclerg	.587	.644	.057	.885	3	**natcity**	.394	.472	.078	.811	3
consci	.505	.562	.057	.955	3	**natfarey**	.650	.728	.078	.848	3
racdif3	.651	.709	.057	.876	3	**localnum**	.721	.799	.078	.854	3
pillok	.559	.617	.058	.887	3	**eqwlth**	.555	.633	.078	.854	3
permoral	.353	.411	.058	.789	1	**natchld**	.527	.606	.079	.824	3
libhomo	.673	.731	.058	.975	3	**natarms**	.614	.693	.079	.841	3
sibs	.838	.897	.058	.910	3	**news**	.745	.825	.080	.841	3
attend	.808	.867	.059	.886	3	**natarmsy**	.582	.663	.081	.823	3
natspac	.674	.734	.059	.888	3	**helpsick**	.545	.627	.082	.829	3
conjudge	.545	.605	.060	.813	3	**teensex**	.602	.684	.082	.865	3
health	.720	.780	.060	.845	3	**suicide2**	.764	.846	.082	.836	3
helpoth	.388	.449	.061	.848	3	**libcom**	.683	.765	.083	.845	3
meovrwrk	.402	.464	.062	.849	3	**satfin**	.642	.725	.083	.789	3
fejobaff	.575	.636	.062	.896	3	**colmil**	.589	.672	.083	.875	3
nataid	.610	.673	.063	.802	3	**earnrs**	.637	.721	.084	.810	3
workhard	.378	.441	.063	.844	3	**socbar**	.780	.865	.085	.838	3
intlwhts	.243	.307	.063	.672	3	**natmass**	.520	.605	.085	.805	3
wlthwhts	.315	.379	.064	.728	3	**xmarsex**	.621	.706	.085	.874	3
conlegis	.528	.593	.065	.868	3	**sppres80**	.697	.782	.085	.853	1
helppoor	.516	.581	.066	.916	3	**chldidel**	.642	.728	.086	.884	3
spkhomo	.759	.825	.066	.859	3	**thnkself**	.510	.596	.086	.824	3

Var	TRT	Heise	Diff	Stability	Nr. panels	Var	TRT	Heise	Diff	Stability	Nr. panels
natsoc	.552	.639	.088	.850	3	**polmurdr**	.497	.606	.108	.779	3
colrac	.534	.622	.088	.872	3	**natroad**	.473	.584	.111	.791	3
socommun	.494	.583	.088	.772	3	**inequal5**	.334	.452	.119	.758	1
nateduc	.623	.712	.089	.838	3	**popular**	.488	.608	.120	.737	3
hapmar	.706	.795	.089	.839	3	**polabuse**	.469	.588	.120	.804	3
spkath	.705	.795	.090	.842	3	**natenviy**	.622	.746	.124	.810	3
rotapple	.430	.521	.091	.700	1	**joblose**	.450	.575	.125	.648	3
satjob	.501	.594	.094	.734	3	**sphrs1**	.564	.692	.128	.683	3
livewhts	.231	.328	.097	.693	3	**natdrugy**	.552	.683	.131	.729	3
conpress	.532	.629	.097	.781	3	**intlblks**	.238	.377	.139	.580	3
natfare	.616	.715	.098	.827	3	**blkwhite**	.512	.654	.142	.650	1
spkrac	.648	.747	.099	.782	3	**contv**	.489	.642	.153	.669	3
tvhours	.618	.717	.099	.777	3	**natheal**	.500	.656	.156	.675	3
nateducy	.664	.766	.102	.817	3	**racwork**	.673	.832	.159	.691	3
natrace	.658	.761	.103	.801	3	**workwhts**	.326	.491	.165	.601	3
natenvir	.644	.749	.105	.794	3	**natcrime**	.492	.661	.169	.601	3
libmil	.595	.700	.105	.803	3	**weekswrk**	.728	.898	.171	.703	1
confinan	.485	.592	.107	.707	3	**finalter**	.401	.580	.179	.531	3
goodlife	.417	.524	.107	.739	3						

Notes: Sample: non-redundant, self- and proxy reports only; excluding performance triads, excluding interviewer and organization reports. TRT, Heise, stability and difference estimates are averaged over common items in the pool.