The Reliability of Survey Measures
RESULTS Series

COMPARISON OF HEISE 3-WAVE SIMPLEX MODEL WITH MULTI-LEVEL MODEL ESTIMATES

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 Model with Multi-level Model Estimates. *The Reliability of Survey Measures Results Series.*

There[1] are several approaches to estimating the amount of measurement error variance that rely on all longitudinal measurement. One of these is the classic multi-level (ML) model, in which the design involves occasions of measurement conceptualized as nested within persons (see Goldstein, 1995). By focusing on within person variation relative to the total variance, this approach provides an estimate of reliability ranging from zero to 1.0, comparable to the SEM simplex approach (Heise, 1969). Both models assume the independence of measurement errors. The multilevel (ML) model has certain intuitive appeal, as it may be viewed as making less restrictive assumptions as the simplex model. Of course, the only assumption made the SEM-simplex approach (hereafter referred to as "the Heise model") is the assumption of the independence of errors and that the correlational data conform to a simplex structure. On the other hand, the ML approach assumes there is no change in the underlying trait being measured.

The ML model can explicitly control for clustering on year, panel and unobserved heterogeneity. The main limitation of the approach is that since it does not endeavor to separate unreliability from true change, and it therefore may tend to underestimate reliability, especially for non-fixed traits of individuals or measures of traits that may change over time. Using this approach, Hout and Hastings (2016) employed a cleverly-constructed mixed-effects model with years as fixed effects (to control for aggregate change over time, net of question unreliability) within such a multilevel framework and estimating the individual level reliability as a random effect. In such a model, unreliability of measurement is defined as any within person change net of wider time effects. As noted, both models involve the same set of assumptions of the measurement model, namely that the errors of measurement are independent.

[1] This document is part of the larger report from this project, co-authored by Duane Alwin and Paula Tufiş. Table and figure numbers refer to those in the source.

This Study

Here we report reliability estimates for approximately 600 measures in the three GSS panels using both methods. Our analysis parallels that of Hout and Hastings (2016) who performed a similar comparison of the two approaches to reliability estimation. We limit our analysis to only non-redundant, self-and proxy reports, excluding performance measures, as well as eliminating interviewer and organization reports. In our Appendix table we present a summary of our two sets of findings for each distinct question in the pool of GSS items considered here, averaged over common items in the pool.

In addition, in that document 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 Heise (1969, eq. 12, page 97)]. [2] These 4-year stability estimates range from high levels, i.e. 1.0 or close to 1.0, to relatively lower levels. As we will report below, the lower the stability of the underlying trait, the greater is the difference between the SEM simplex and multi-level approaches to estimating reliability.

Results

The table presented in the appendix provides the detailed comparisons of the two approaches. We present a summary of these results in Table 1. In general, as expected the Heise 3-wave simplex model estimates are greater than the multi-level model estimates, although there are a substantial

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

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 trais that theoretically or practically 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.

Table 1. Heise and multilevel reliability ($\hat{\rho}$) estimates and differences by stability, averaged over GSS panels, for non-redundant self- and proxy-reports

	Number of measures	Stability	Heise	$\hat{\rho}$	Diff	t-test	df	p-value
Fixed traits	11	0.975	.872	.858	.014	1.841	10	.095
Highly stable traits (stability = .93 - 1.0)	53	0.963	.766	.751	.015	5.209	52	.000
Relatively stable traits (stability = .87 - .92)	53	0.902	.717	.665	.052	15.510	52	.000
Less stable traits (stability = .82 - .86)	53	0.844	.661	.592	.069	18.607	52	.000
Unstable traits (stability < .82)	52	0.743	.605	.501	.104	17.061	51	.000

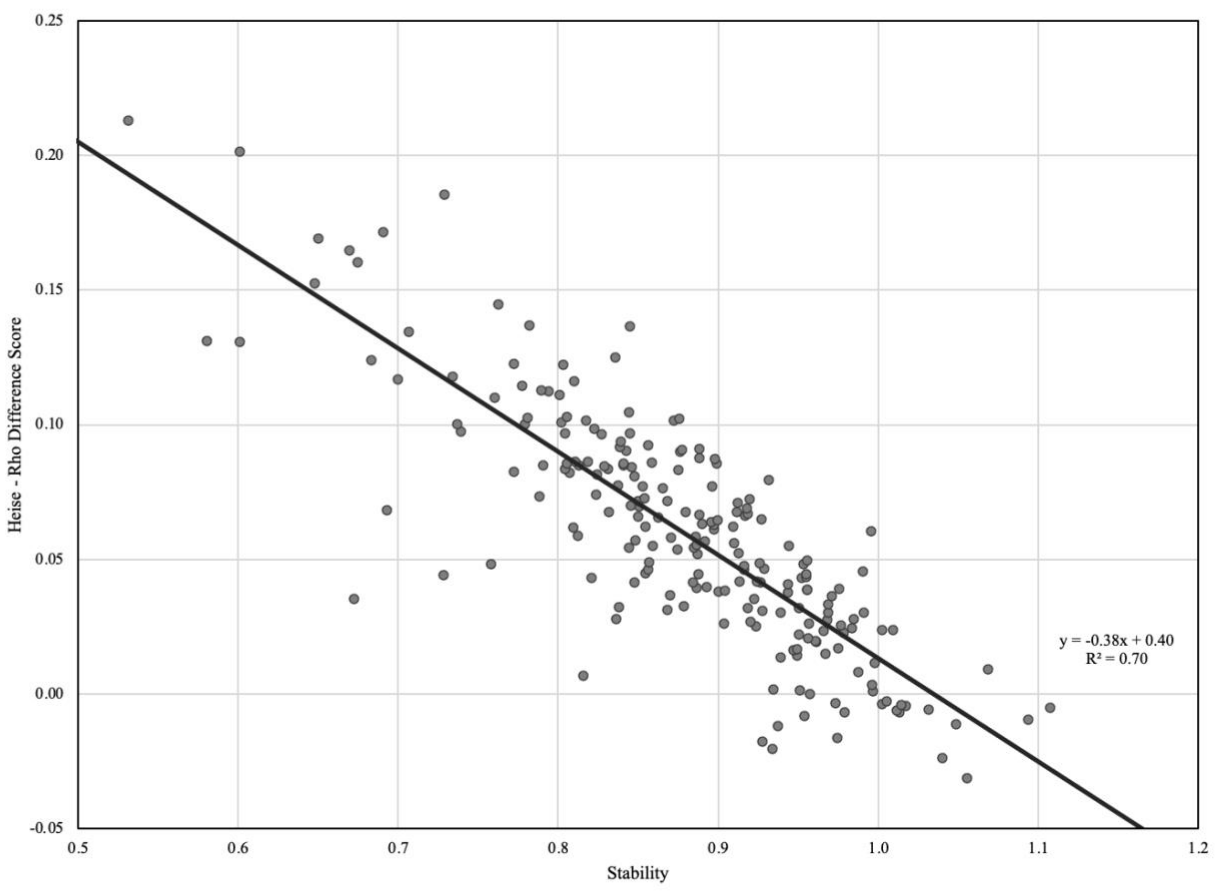
As indicated in this summary table (for more detail see the appendix), we performed a test of the difference between the ML estimate (denoted $\hat{\rho}$ in the table of results) and the Heise estimates, using a test of "matched pairs" (see Blalock 1972, pp. 233-235). These results indicate that for "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, the differences are greater and statistically significant.

[3] There was a small number of cases (19 cases of 211) where the ML estimates were greater than the Heise estimates.

Stability of Latent Traits

As predicted by Hout and Hastings (2016), there is a strong relationship between the differences in these two reliability estimates and the fixed nature and/or the stability of the underlying trait being measured. These patterns are depicted in Figure 1, where we present a scatterplot relating the difference score [i.e., the Heise minus the ML estimates] to the level of stability, and the linear regression of the difference on stability ($R^2 = 0.70$). The results summarized here clearly suggest that the difference between the two estimate is a relatively linear function of the stability of the trait being measured. Not surprisingly, Kiley and Vaisey's (2021) results anticipate the fact that many of the GSS items reveal high levels of stability over the four-year period.

Figure 1. Scatterplot of the relationship between the Heise-$\hat{\rho}$ difference score and the level of stability in the underlying trait



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 the question, that is, what the trait involves. Content can be factual (i.e., objective information that can be easily verified) or non-factual, or subjective, in nature. Non-factual content can be further classified as traits involving beliefs, values, attitudes, expectations, or self-perceptions/evaluations. There is a well-established finding in the survey methods literature that the measurement of factual content (e.g., birth year) can be assessed more accurately in surveys than non-facts in survey reports (e.g., Alwin, 2007). Thus, we hypothesized that the content being measured may be related to the differences between the two approaches to reliability estimation.

To examine this hypothesis, we present the mean estimates of reliability for self- and proxy-reports, averaged across the three GSS panels, organized by question content and the approach to reliability estimation. This table permits us to analyze the differences between the ML and Heise estimates within categories of content. Question content is operationalized here according to Alwin's (2007, pages 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, specifically, beliefs (statements about what is), attitudes (positive and negative sentiments toward a social object, values (statements about what should be), self-perceptions (beliefs about the self), self-assessments (evaluations of the self) and expectations (beliefs about future events or situations).

Table 2. Heise and multilevel reliability ($\hat{\rho}$) estimates, by question content and approach to reliability estimation, averaged across GSS panels, for non-redundant self- and proxy-reports

				Heise - $\hat{\rho}$ Comparisons		
Content	Measures	Heise	$\hat{\rho}$	t test	df	p-value
Facts	35	.847	.797	6.112	34	.000
Non-facts	176	.656	.594	19.002	175	.000
Beliefs	67	.634	.564	12.819	66	.000
Values	42	.670	.614	9.239	41	.000
Attitudes	35	.671	.614	9.478	34	.000
Self-Assessments	12	.652	.576	4.429	11	.001
Self-Perceptions	14	.740	.701	4.344	13	.001
Expectations	6	.532	.465	2.627	5	.047
Total	211	.688	.628	19.685	210	.000
Comparisons						
All content						
F-ratio		13.073	14.797			
p-value		.000	.000			
Facts vs. Non-facts						
F-ratio		61.118	63.843			
p-value		.000	.000			
Within Nonfacts						
F-ratio		2.410	3.593			
p-value		.039	.004			

The results in Table 2 provide a formal test of the differences with categories of content, as noted, facts vs. non-facts, and subcategories of non-facts. We employ, as above, the "paired samples" t-test procedure, which compares the means of two variables for a single group (see Blalock, 1972). This procedure tests whether the differences in the two approaches to reliability estimation differ from 0.00. The results in this table consistently reveal systematic differences between them, with the Heise estimates averaging at higher levels compared to the ML approach.

Consistent with prior research, these results also demonstrate that questions assessing subjective content have significantly lower reliabilities (see Alwin, 2007, pp. 158-162). Among subjective categories of content, self-perceptions have the highest levels of reliability. Expectations are measured with the least reliability. Both approaches to reliability estimation reveal these patterns.

Stability vs. Content

We further examine the relationship between stability and reliability estimates using linear regression as a way of summarizing the observed patterns. Table 3 presents a series of regression models that parameterize the effects of several predictor variables on the difference between the two estimates (Heise minus ML reliability estimates).

Table 3. Regression of differences in Heise and multilevel reliability estimates on attributes of questions: pooled GSS panels

Predictors	Model [1] 1		2		3		4		5	
Intercept	.704	***	.067	***	.144	***	.064	***	.070	***
ρ (centered)	.899	***	----		----		----		----	
Stability (centered) [2]			-.043	***	----		----		-.044	***
Stability quartiles [3]										
2nd quartile					-.065	***	----		----	
3rd quartile					-.100	***	----		----	
4th quartile					-.129	***	----		----	
Content: fact versus non-fact [4]										
Non-facts--beliefs							0.021	**	-0.005	
Non-facts--values							0.012		0.002	
Non-facts--attitudes							0.009		-0.001	
Non-facts--self assessments							0.025	*	-0.010	+
Non-facts--self perceptions							-0.009		-0.013	**
Non-facts--expectations							0.030	+	-0.010	
R^2	0.840		0.794		0.651		0.046		0.799	
N of cases	598		598		598		598		598	

*Key: + p<0.10, * p<0.05, ** p<0.01, *** p<0.001*
[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 ρ-reliability
Model 2: Regress Heise-ρ Difference on Stability (centered)
Model 3: Regress Heise-ρ Difference on Stability as quartiles
Model 4: Regress Heise-ρ Difference on Facts vs. type of non-facts
Model 5: Regress Heise-ρ 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-ρ Difference score.
Note: In Model 4 and 5 "facts" is the omitted category

The first model in this table of regressions reveals the convergences between the two estimates of reliability. The linear relationship between the two estimates is high (R^2 = 0.840), but this does not mean they are identical; see, e.g., the bivariate scatterplot in Figure 1. The remaining models in Table 3 regress the difference (i.e., Heise – ML) on the factors considered earlier, stability and content. As revealed in model 2 of Table 3, the difference is highly predictable from the 4-year stability estimate (R^2 = 0.794). This model parameterizes the linear relationship previously reported in Figure 1 above. In model 3 we regress the difference on stability using quartiles as dummy variables, reinforcing the finding that the relationship is linear.

In model 4 of Table 3, we regress the difference between the two reliability 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 model is the facts category. These results indicate there is a significant difference between facts and three types of non-facts, specifically, beliefs, self-assessments, and expectations. All other types of non-facts are not significantly different from facts with respect to the differences between the two reliability estimates.

Finally, in model 5, the difference score is regressed on the dummy variables representing content categories, while controlling for the stability of the underlying trait measured by the question. These results also 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 effects are removed from these contrasts.

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 NY: McGraw-Hill.

Goldstein, H. 1995. *Multilevel Statistical Models*. 2nd edition. New York NY: Halsted Press.

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

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

Kiley, K., and S. Vaisey. 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	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels	Var	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels
letin1	.588	.557	-.031	1.056	3	**natcityy**	.470	.495	.025	.983	3
degree	.911	.887	-.024	1.040	3	**absingle**	.833	.859	.026	.977	3
incom16	.615	.594	-.021	.934	3	**bible**	.755	.781	.026	.904	3
finrela	.628	.611	-.017	.928	3	**speduc**	.896	.922	.026	.957	3
childs	.927	.911	-.016	.975	3	**gunlaw**	.649	.676	.027	.984	3
workblks	.377	.365	-.012	.937	3	**helpnot**	.482	.509	.027	.920	3
coneduc	.491	.480	-.011	1.048	3	**relactiv**	.677	.705	.028	.836	3
inequal3	.449	.440	-.009	1.094	1	**spdeg**	.907	.935	.028	.968	3
postlife	.925	.917	-.008	.954	3	**fepol**	.666	.696	.030	.991	3
spfund	.862	.855	-.007	.979	3	**abany**	.822	.852	.030	.969	3
agekdbrn	.940	.933	-.007	1.013	3	**abpoor**	.851	.881	.030	.939	3
maeduc	.883	.877	-.006	1.011	3	**abrape**	.880	.910	.030	.928	3
fehire	.460	.454	-.006	1.032	3	**socfrend**	.510	.541	.031	.868	3
librac	.543	.538	-.005	1.107	3	**relpersn**	.794	.826	.032	.919	3
mapres80	.775	.770	-.005	1.015	1	**discaffm**	.333	.365	.032	.838	3
natracey	.657	.652	-.005	1.017	3	**nataidy**	.633	.665	.032	.950	3
liveblks	.418	.414	-.004	1.002	3	**closewht**	.466	.499	.033	.878	3
divlaw2	.847	.844	-.003	1.005	3	**natdrug**	.438	.472	.034	.969	3
pornlaw	.633	.630	-.003	.973	3	**spanking**	.665	.700	.035	.922	3
class	.702	.702	.000	.957	3	**intlwhts**	.272	.307	.035	.672	3
cohort	.994	.995	.001	.996	3	**libath**	.607	.643	.036	.971	3
discaffw	.407	.408	.001	.951	3	**marwht**	.379	.416	.037	.870	3
polviews	.669	.670	.001	.934	3	**abnomore**	.833	.871	.038	.944	3
polattak	.543	.546	.003	.996	3	**natpark**	.472	.510	.038	.900	3
wlthblks	.330	.337	.007	.816	3	**popespks**	.582	.620	.038	.904	3
discaff	.399	.408	.009	1.068	3	**marblk**	.602	.641	.039	.887	3
paeduc	.922	.931	.009	.988	3	**fefam**	.612	.651	.039	.893	3
colhomo	.754	.766	.012	.998	3	**libhomo**	.692	.731	.039	.975	3
fepresch	.555	.569	.014	.949	3	**premarsx**	.773	.812	.039	.956	3
fund16	.856	.870	.014	.939	3	**racdif2**	.640	.679	.039	.956	3
incgap	.453	.468	.015	.967	1	**pray**	.812	.853	.041	.926	3
fund	.860	.876	.016	.949	3	**getahead**	.435	.476	.041	.944	3
educ	.897	.914	.017	.975	3	**helpoth**	.408	.449	.041	.848	3
god	.829	.846	.017	.947	3	**chldidel**	.686	.728	.042	.884	3
abdefect	.841	.860	.019	.961	3	**papres80**	.752	.794	.042	.924	1
rellife	.664	.683	.019	.961	1	**partyid2**	.868	.910	.042	.913	3
reborn	.903	.924	.021	.956	3	**trust2**	.788	.831	.043	.955	3
life	.632	.654	.022	.950	3	**homosex**	.861	.904	.043	.952	3
padeg	.917	.940	.023	1.009	3	**socrel**	.543	.587	.044	.821	3

Var	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels	Var	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels
polhitok	.737	.760	.023	.978	3	**wlthwhts**	.335	.379	.044	.728	3
suicide1	.773	.797	.024	1.003	3	**consci**	.518	.562	.044	.955	3
helpblk	.603	.627	.024	.966	3	**madeg**	.877	.922	.045	.990	3
conbus	.504	.529	.025	.923	3	**marasian**	.490	.535	.045	.855	3
parsol	.612	.657	.045	.888	3	**conarmy**	.544	.614	.070	.851	3
helppoor	.535	.581	.046	.916	3	**spkmil**	.626	.696	.070	.912	3
closeblk	.616	.662	.046	.856	3	**health**	.710	.780	.070	.845	3
spkcom	.771	.818	.047	.929	3	**natsoc**	.568	.639	.071	.850	3
sexeduc	.779	.827	.048	.953	3	**conlegis**	.521	.593	.072	.868	3
inequal5	.404	.452	.048	.758	1	**prayer**	.689	.761	.072	.920	3
cappun	.838	.886	.048	.926	3	**eqwlth**	.560	.633	.073	.854	3
wrkwayup	.584	.632	.048	.916	3	**permoral**	.338	.411	.073	.789	1
marhisp	.497	.546	.049	.857	3	**thnkself**	.522	.596	.074	.824	3
colath	.631	.681	.050	.956	3	**teensex**	.608	.684	.076	.865	3
grass	.859	.911	.052	.913	3	**socbar**	.788	.865	.077	.838	3
pillok	.565	.617	.052	.887	3	**sppres80**	.705	.782	.077	.853	1
xmarsex	.652	.706	.054	.874	3	**racopen2**	.580	.657	.077	.896	3
workhard	.387	.441	.054	.844	3	**abhlth**	.808	.887	.079	.931	3
conclerg	.590	.644	.054	.885	3	**natchld**	.525	.606	.081	.824	3
attend	.812	.867	.055	.886	3	**natfarey**	.647	.728	.081	.848	3
conlabor	.533	.588	.055	.859	3	**natsci**	.464	.546	.082	.807	3
helpful2	.681	.736	.055	.944	3	**socommun**	.500	.583	.083	.772	3
sibs	.841	.897	.056	.910	3	**raclive**	.767	.850	.083	.875	3
meovrwrk	.407	.464	.057	.849	3	**conmedic**	.471	.554	.083	.804	3
punsin	.574	.631	.057	.891	1	**natcrimy**	.587	.670	.083	.831	3
obey	.606	.664	.058	.871	3	**news**	.741	.825	.084	.841	3
hrs1	.528	.587	.059	.812	3	**prestg80**	.690	.774	.084	.846	1
sprtprsn	.741	.800	.059	.886	3	**conjudge**	.520	.605	.085	.813	3
granborn	.907	.968	.061	.995	3	**natroad**	.499	.584	.085	.791	3
letdie1	.762	.823	.061	.897	3	**helpsick**	.542	.627	.085	.829	3
earnrs	.659	.721	.062	.810	3	**uswary**	.653	.738	.085	.899	3
localnum	.737	.799	.062	.854	3	**natarms**	.607	.693	.086	.841	3
suicide4	.742	.804	.062	.910	3	**natcity**	.386	.472	.086	.811	3
nathealy	.511	.574	.063	.890	3	**natmass**	.519	.605	.086	.805	3
xmovie	.794	.857	.063	.897	3	**rincom06**	.731	.817	.086	.819	3
fejobaff	.572	.636	.064	.896	3	**spkhomo**	.739	.825	.086	.859	3
marhomo	.771	.836	.065	.900	3	**aged2**	.637	.724	.087	.898	3
fair2	.734	.799	.065	.927	3	**suicide3**	.731	.818	.087	.888	3
tax	.615	.680	.065	.862	3	**courts2**	.771	.861	.090	.877	3
fechld	.530	.596	.066	.850	3	**racdif3**	.619	.709	.090	.876	3
racdif1	.652	.718	.066	.917	3	**spkath**	.705	.795	.090	.842	3

Var	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels	Var	$\hat{\rho}$	Heise	Diff.	Stability	Nr. panels
natspac	.667	.734	.067	.888	3	**racdif4**	.608	.699	.091	.888	3
polescap	.542	.609	.067	.912	3	**nateduc**	.620	.712	.092	.838	3
reliten2	.849	.916	.067	.919	3	**richwork**	.666	.759	.093	.856	3
fear	.684	.752	.068	.918	3	**hapmar**	.702	.795	.093	.839	3
happy	.524	.592	.068	.832	3	**polabuse**	.492	.588	.096	.804	3
livewhts	.260	.328	.068	.693	3	**natfare**	.618	.715	.097	.827	3
affrmact	.578	.646	.068	.879	3	**libcom**	.668	.765	.097	.845	3
goodlife	.427	.524	.097	.739	3	**libmil**	.578	.700	.122	.803	3
natarmsy	.565	.663	.098	.823	3	**jobfind**	.574	.697	.123	.773	3
popular	.508	.608	.100	.737	3	**sphrs1**	.568	.692	.124	.683	3
colrac	.521	.622	.101	.872	3	**suicide2**	.721	.846	.125	.836	3
polmurdr	.505	.606	.101	.779	3	**workwhts**	.361	.491	.130	.601	3
nataid	.572	.673	.101	.802	3	**intlblks**	.246	.377	.131	.580	3
nateducy	.664	.766	.102	.817	3	**confinan**	.457	.592	.135	.707	3
colmil	.570	.672	.102	.875	3	**spkrac**	.610	.747	.137	.782	3
conpress	.526	.629	.103	.781	3	**income06**	.744	.881	.137	.845	3
natspacy	.632	.735	.103	.806	3	**weekswrk**	.728	.873	.145	.762	3
colcom	.591	.696	.105	.844	3	**joblose**	.423	.575	.152	.648	3
kidssol	.569	.679	.110	.761	3	**natheal**	.495	.656	.161	.675	3
natrace	.650	.761	.111	.801	3	**contv**	.477	.642	.165	.669	3
natenvir	.636	.749	.113	.794	3	**blkwhite**	.485	.654	.169	.650	1
satfin	.612	.725	.113	.789	3	**racwork**	.661	.832	.171	.691	3
tvhours	.603	.717	.114	.777	3	**natdrugy**	.497	.683	.186	.729	3
natenviy	.630	.746	.116	.810	3	**natcrime**	.460	.661	.201	.601	3
rotapple	.404	.521	.117	.700	1	**finalter**	.367	.580	.213	.531	3
satjob	.477	.594	.117	.734	3						

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