Registered Report

Structural Validity Evidence for the Oxford Utilitarianism Scale Across 15 Languages

Briana Oshiro[1], William H. B. McAuliffe[2], Raymond Luong[1], Anabela C. Santos[3,4], Andrej Findor[5], Anna O. Kuzminska[6], Anthony Lantian[7], Asil A. Özdoğru[8], Balazs Aczel[9], Bojana M. Dinić[10], Christopher R. Chartier[11], Jasper Hidding[12], Job A. M. de Grefte[13], John Protzko[14], Mairead Shaw[1], Maximilian A. Primbs[15], Nicholas A. Coles[16], Patricia Arriaga[17], Patrick S. Forscher[18], Savannah C. Lewis[11], Tamás Nagy[9], Wieteke C. de Vries[12], William Jimenez-Leal[19], Yansong Li[20], and Jessica Kay Flake[1]

[1]Department of Psychology, McGill University, Montreal, QC, Canada
[2]Division on Addiction, Cambridge Health Alliance, Malden, MA, USA
[3]Aventura Social and Instituto de Saúde Ambiental (ISAMB), Faculdade de Medicina, Universidade de Lisboa, Portugal
[4]CIS-IUL, ISCTE – Instituto Universitário de Lisboa, Portugal
[5]Faculty of Social and Economic Sciences, Comenius University in Bratislava, Slovakia
[6]Faculty of Management, University of Warsaw, Poland
[7]Department of Psychology, Université Paris Nanterre, France
[8]Department of Psychology, Marmara University, Istanbul, Turkey
[9]Institute of Psychology, Eotvos Lorand University, Budapest, Hungary
[10]Department of Psychology, Faculty of Philosophy, University of Novi Sad, Serbia
[11]Department of Psychology, Ashland University, OH, USA
[12]Department of Marketing, University of Groningen, The Netherlands
[13]Department of Philosophy, University of Groningen, The Netherlands
[14]Department of Psychological Science, Central Connecticut State University, New Britain, CT, USA
[15]Behavioral Science Institute, Radboud University, The Netherlands
[16]Center for the Study of Language and Information, Stanford University, CA, USA
[17]ISCTE – Instituto Universitá rio de Lisboa (IUL), Portugal
[18]Interuniversity Laboratory of Psychology, Personality, Cognition, Social Change (LIP/PC25), Université Grenoble Alpes, France
[19]Department of Psychology, Universidad de los Andes, Bogota, Colombia
[20]Department of Psychology, Nanjing University, Nanjing, People's Republic of China

Abstract: *Background*: The Psychological Science Accelerator (PSA) recently completed a large-scale moral psychology study using translated versions of the Oxford Utilitarianism Scale (OUS). However, the translated versions have no validity evidence. *Objective*: The study investigated the structural validity evidence of the OUS across 15 translated versions and produced version-specific validity reports. *Methods*: We analyzed OUS data from the PSA, which was collected internationally on a centralized online questionnaire. We also collected qualitative feedback from experts for each translated version. *Results*: For each version, we produced version-specific psychometric reports which include the following: (1) descriptive item and demographics analyses, (2) factor structure evidence using confirmatory factor analyses, (3) measurement invariance testing across languages using multiple-group confirmatory factor analyses and alignment optimization, and (4) reliability analyses using coefficients α and ω.

Keywords: Oxford Utilitarianism Scale, translation, measurement invariance, reliability, Psychological Science Accelerator

The Psychological Science Accelerator (PSA) is a big team science collaborative dedicated to large multicultural and multilingual studies with more than 1,200 members from 71 countries as of 2020 (Paris et al., 2020). The PSA completed a study of moral reasoning, PSA 006 (Bago et al., 2022), which used translated versions of the Oxford Utilitarianism Scale (OUS; Kahane et al., 2018). We performed a structural validation study of 15 translated versions of the OUS using the PSA 006 data. Validation of the translated OUS versions and extending beyond its initial development, which was limited to MTurk workers, professional philosophers (Kahane et al., 2018), and Turkish university students (Erzi, 2019), will facilitate the study of moral cognition across all human societies, as is generally intended (e.g., Awad et al., 2020; Mikhail, 2007). This requires, at minimum, gathering validity evidence from a large and diverse sample taken from multiple populations (Henrich et al., 2010).

No validity evidence for the translated versions created by the PSA has been gathered yet. Any use of the OUS in populations other than those on which it has been validated, such as comparing scores across translated versions of the OUS to rank populations in terms of their support for utilitarianism, requires evidence of measurement invariance. Without such evidence, results from downstream analyses can become confounded with differences in what the OUS measures across populations (Chen, 2008; Guenole & Brown, 2014; Slaney & Maraun, 2008).

Our large-scale study of the psychometric properties of the OUS translations includes descriptive item analysis, qualitative feedback from expert translators, confirmatory factor analyses, measurement invariance testing, and reliability reporting. Individual psychometric validity reports, which detail the complete findings for each specific translated version, are provided as supplementary materials. To ground the proposed work, we first review the theory, development, and existing validity evidence for the OUS.

Measuring Utilitarianism

Utilitarianism is the view that a person's behavior is morally good to the extent that it produces a greater amount of net well-being than any other available action (for a primer, see Smart & Williams, 1973). In moral philosophy, utilitarianism is an exemplar of *consequentialist* philosophies that judge the value of actions by their consequences. It stands in contrast to *deontological* theories that judge actions by their actors' intentions to uphold values that have worth over and above their total effects on well-being (e.g., equity, esthetic beauty, etc.) and *virtue* theories that judge actions as good to the extent that they reflect a wise deployment of positive character traits (e.g., honesty, loyalty).

The OUS (Kahane et al., 2018) is a questionnaire developed to overcome the limitations of previous instruments that use responses to hypothetical vignettes to measure endorsement of utilitarianism. For example, the most popular type vignettes, *trolley dilemmas*, ask participants to indicate whether they would be willing to intentionally put one person in front of a fast-moving trolley to prevent it from running over a larger number of people. Research suggests that these vignettes mostly capture variation in aversion to "instrumental harm," or causing innocent people harm to increase the overall amount of well-being (Everett & Kahane, 2020). For instance, utilitarianism would require oppressing a minority group if doing so would make the majority group much better off, as the improved welfare of the latter outweighs the reduced welfare of the former. The only utilitarian case against minority oppression would be that it might have unintended costs that outweigh the increased happiness of the majority group (Smart & Williams, 1973). Although including items about instrumental harm is a necessary aspect of a utilitarianism measure, it is not sufficient because endorsement of instrumental harm can result not only from reflecting on utilitarianism's premises but also from traits such as sadism. Furthermore, in practice, most prominent utilitarians give some weight to deontological considerations and reject instrumental harm due to its negative side effects (e.g., Todd & MacAskill, 2017). What instead makes utilitarians distinctive is their emphasis on "impartial beneficence," or the obligation to benefit others in the most impartial, efficient manner possible.

The OUS provides comprehensive construct coverage by explicitly acknowledging the distinctness of obligations to harm and obligations to help using an instrumental harm subscale and an impartial beneficence subscale. As shown in Table 1, the impartial beneficence subscale comprises five items (e.g., "From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy"); the instrumental harm subscale comprises four (e.g., "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people"). Disagreement with the items does not imply an endorsement of a particular moral theory, only a lack of endorsement of utilitarianism (Kahane et al., 2018, p. 136).

Table 1. The 9-item Oxford Utilitarianism Scale (OUS)

Subscale	Item
Impartial beneficence	IB1. From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.
	IB2. From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.
	IB3. If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.
	IB4. It is just as wrong to fail to help someone as it is to actively harm them yourself.
	IB5. It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.
Instrumental harm	IH1. It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.
	IH2. If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.
	IH3. It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.
	IH4. Sometimes it is morally necessary for innocent people to die as collateral damage – if more people are saved overall.

Note. All items are rated on 7-point Likert scale (1 = *strongly disagree*; 4 = *neither agree nor disagree*; 7 = *strongly agree*). From Kahane et al., 2018.

OUS Scale Development

Item Generation

Kahane et al. (2018) constructed the initial pool of over 200 items from existing studies that measured utilitarian judgment. They refined this pool by consulting theory on moral philosophy and expert moral philosophers on utilitarianism. The authors then examined each item and filtered out items that were redundant, irrelevant, poorly worded, or confusing – resulting in 94 items. Finally, the authors asked a panel of professional moral philosophers to rate the quality of each of the 94 refined items on 5-point scales (e.g., "How good do you think this item is for discriminating utilitarian and nonutilitarian views?") and provide qualitative feedback on the items. Based on this quantitative and qualitative feedback, the authors reached a final pool of 77 items. The key requirement for item content was that the final set of items captures not only the negative aspect of utilitarianism – namely that it does not disallow any particular action, so long as it brings about the most benefits – but also the positive aspect of utilitarianism – namely that it mandates the maximization of welfare.

Structural Validation

The developers of the OUS conducted psychometric analyses across numerous samples to determine the scale's factor structure. First, they conducted exploratory factor analyses on a sample of 960 MTurk workers (Study 1; Kahane et al., 2018), initially suggesting a four-factor structure. These four factors included, in addition to impartial beneficence and instrumental harm, "anti-traditional morality," which refers to "deontological ideas associated with conservative thought" (p. 143, e.g., "Criminals should receive the punishment they deserve – even if this will not protect the public or deter crime in the future), and "truth-telling and promise-keeping," which contained items pertaining to honesty and keeping promises (e.g., "It is morally permissible to lie if doing so would help others a great deal"). Antitraditional morality and truth-telling and promise-keeping were dropped, however, when confirmatory factor analyses (CFAs) on the same sample (Study 1) and on a subsequent independent sample of 282 MTurk workers (Study 2) supported a two-factor solution that only included impartial beneficence and instrumental harm. This two-factor structure was again supported in a subsequent CFA with a Turkish university student sample (Erzi, 2019). To date, measurement invariance has not been formally evaluated.

Convergent Validity

After supporting the factor structure of the scale, the developers of the OUS then conducted correlation analyses between the two subscales, the total scale, and other related measures of utilitarianism (Table 2; reproduced from Kahane et al., 2018, p. 146). These related measures included explicit utilitarianism and moral dilemmas (e.g., trolley problems). Both subscales were found to be strongly correlated with the total score, while the subscales were expected to be only weakly correlated to each other. Although utilitarianism requires embracing both impartial concern and instrumental harm, the scale developers argued they correlate only weakly among laypeople because different psychological traits predispose individuals to endorse impartial concern and instrumental harm (Kahane et al., 2018). As expected, the subscales and total scale also correlated with the other related measures of utilitarianism.

Table 2. Correlations between the OUS and other measures of utilitarianism

Measure	1	2	3
1. Overall Oxford Utilitarianism Scale	—		
2. Impartial Beneficence Subscale (OUS-IB)	.81**	—	
3. Instrumental Harm Subscale (OUS-IH)	.70**	.14*	—
4. Explicit utilitarianism	.35**	.37*	.13*
5. Classic sacrificial dilemmas	−.34**	−.21*	−.32**
6. Greater good dilemmas	.40**	.50**	.07**

Note. From Kahane et al., 2018.
*$p < .01$. **$p < .005$.

Intended Use

Kahane et al. (2018) aimed to demonstrate the OUS's utility for moral psychology research. Researchers have generally been interested in measures of utilitarianism to test to what extent *commonsense morality*, widely shared moral judgments among nonphilosophers, deviates from what utilitarianism would prescribe (Everett & Kahane, 2020). For example, many people would agree that instrumental harm is worse than harming others as a foreseen side effect of an otherwise beneficial act, although this distinction is irrelevant in the utilitarian framework (Royzman & Baron, 2002). Such deviations have formed the basis for several theories of moral cognition, including a *moral grammar* that underpins legal systems (Mikhail, 2007), dual process theories that emphasize emotional processes (Greene & Haidt, 2002), and deontological theories of *sacred values* (Tetlock, 2003). The OUS differs from the measures of utilitarianism used in these studies in that it is first and foremost a measure of individual differences designed to have some cross-situational consistency. It is most well-suited to "study the relationship between utilitarian tendencies and various other traits to advance our understanding of proto-utilitarian thinking [i.e., utilitarian judgments among those without exposure to utilitarian philosophy]" (Kahane et al., 2018; p. 138), although the OUS can also be used to inform the factors that might have influenced specific moral judgments (p. 138).

Target Population

Although the idea that an action's consequences are relevant to its moral value is widespread, utilitarianism's assertion that the value of an action should be judged solely by whether it achieved the best consequences possible is not common in public discourse (Smart & Williams, 1973). This makes measuring endorsement of utilitarianism among laypeople practically difficult because they have not had an opportunity to recognize and reflect on its implications. Even people who weigh consequences in their moral judgments more strongly than most may balk at utilitarianism if they were told it implies that showing partiality to family over strangers is immoral (Law et al., 2022). Conversely, people who initially disagree with the controversial implications of utilitarianism might change their mind if they were exposed to arguments in favor of utilitarianism that are common in the moral philosophy literature. Also, familiarity with the core tenets of utilitarianism may be necessary to consistently applying utilitarian reasoning to a wide range of issues.

The OUS circumvents these issues by measuring whether people agree with some of utilitarianism's distinctive implications. This approach is premised on the assumption that people who already agree with unpopular implications of utilitarianism, even before they learn about it, would be more likely to endorse utilitarianism were they to eventually learn about it (Kahane et al., 2018, p. 138). Thus, the OUS was designed for use with nonphilosophers who have not learned about the arguments offered in favor and against utilitarianism or the theories of right and wrong against which it is normally contrasted. People with formal training in moral philosophy do not require a multi-item questionnaire: "The OUS is not designed specifically for those with substantial experience with the theory of utilitarianism (one does not need a scale to measure such an expert's view – you can just ask them!)" (Kahane et al., 2018, p. 150). We therefore wanted to assess whether the OUS can be used and indeed measures the same underlying beliefs across numerous languages and cultural contexts in nonexpert populations.

Method

All materials, analysis code, and psychometric reports are openly available on the Open Science Framework (https://osf.io/y96rm/).

Design and Analysis Transparency

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. When we use inferential tests, we report exact *p*-values, effect sizes, and 95% confidence or credible intervals.

Data Collection Procedure

This registered report focuses on the psychometric analyses of existing data collected from previously translated instruments. The PSA 006 study (Bago et al., 2022) was administered through a centralized online survey, with all participating labs conforming to their country's and

institution's requirements for ethical data collection (e.g., ethical review and approval, if required). The PSA's translation teams translated and back-translated the two scales from the original English versions to 22 languages (our analysis includes 15, described below). The Turkish scale was not translated but reused from an earlier adaption (Erzi, 2019). The translation process for each language proceeded in five steps (here, the original document refers to the original OUS version and the source language is English):

1. Translation: The original document is translated from source language A to target language B by a first set of translators, resulting in document Version 1.
2. Back-translation: Version 1 is translated back from target language B to source language A independently by a second set of translators, resulting in Version 2.
3. Discussion: Versions 1 and 2 are discussed among the first set of translators, the second set of translators, and the language coordinator. Discrepancies in Versions 1 and 2 are detected and solutions are discussed. Version 3 is created.
4. External Readings: Version 3 is tested on two nonacademics fluent in the target language B. Members of the fluent group are asked how they perceive and understand the translation. Possible misunderstandings are noted and again discussed as in Step 3.
5. Cultural Adjustments: Data collection labs read materials and identify any needed adjustments for their local participant sample. Adjustments are discussed with the language coordinator, who makes any necessary changes, resulting in the final version for each site (Psychological Science Accelerator, 2018).

Participants

Sample Size

Our registered analysis uses multiple-group CFA. To appropriately plan the sample size requirements for analysis, we identified a core set of articles (French & Finch, 2006; Koziol & Bovaird, 2018; Meade & Bauer, 2007; Meade et al., 2008) that contained relevant simulation conditions: multiple group models with multiple factors, three to eight items per factor, and a range of sample sizes feasible to the current work (N = 200–<1,000). Based on this review, we determined that our minimum sample size for a translation to undergo psychometric analysis would be N = 400.

Exclusion Criteria

We excluded any participants who (1) did not finish the survey, (2) provided incorrect answers to any of three careless responding checks (see below), (3) found the survey materials confusing, (4) had technical problems while completing the survey, or (5) did not fill out the survey in their self-reported native language (complete wordings of exclusion questions are included in the supplemental materials for each version). After correspondence with the original authors, we additionally excluded data from practice runs and any data from labs that did not flag practice runs.

Surveys administered in noninterview settings often yield substantial amounts of careless responding (Meade & Craig, 2012). Including responses from inattentive participants can cause model misfit, potentially creating an illusion of misspecification (Arias et al., 2020). Moreover, the rate of careless responding can differ across cultural groups (Grau et al., 2019), which could artificially induce noninvariance across groups. To address these issues, we screened out inattentive participants before fitting models. At the end of the survey, each participant completed three yes-or-no items: (1) I was born on February 30th, (2) I've travelled to the Moon three times, and (3) I can read and write. Although one could correctly answer all these questions randomly, they would always get at least one item incorrect if they answered items in a straight line, a common form of careless responding (Arias et al., 2020). We excluded a participant if they responded to any of the items incorrectly (i.e., "Yes" to Item 1 and Item 2 or "No" to Item 3).

Data collection for the PSA 006 project (Bago et al., 2022) concluded in December 2020. Table 3 reports all sample size differences resulting from our exclusions. Beyond frequency counts for sample sizes of each language, we had not examined the raw data and confirmed that the proposed analyses had not been conducted at the time of Stage 1 submission. Given our realized exclusions, 15 of the 23 versions had enough data to be included in the psychometric analyses.[1]

Data Analysis

For each version of the OUS, we generated a supplementary psychometric report containing detailed information about descriptive statistics and item level analyses, qualitative item feedback from translators, measurement invariance with respect to the original English version, and

[1] Note that the Dutch version of the OUS also appeared to meet the sample size threshold at the time of the Stage 1 submission. However, participants from Dutch labs also had the option to complete the questionnaire in English. We were unable to confirm the exact number of Dutch versions completed without further analysis of the data, so we did not include the Dutch version in the registered report proposal. After analyzing the data during the Stage 2 analysis, we found that the Dutch version did not meet sample size threshold.

Table 3. Sample sizes and exclusions of eligible OUS languages

Language	*N* completed, before exclusions	*N* final, after exclusions (% completed)
US English (reference)	8,624	6,325 (73%)
Chinese (simplified)	1,605	883 (55%)
French	1,390	1,096 (79%)
German	2,976	2,605 (88%)
Greek	531	399 (75%)
Hungarian	942	777 (82%)
Italian	511	401 (78%)
Polish	1,423	1,024 (72%)
Portuguese (Portugal)	750	589 (79%)
Romanian	769	630 (82%)
Russian	580	414 (71%)
Serbian*	549	424 (77%)
Slovak	565	461 (82%)
Spanish	1,137	869 (76%)
Turkish	1,618	1,111 (69%)

Note. Third-party reports of sample size for the Stage 1 manuscript did not indicate that the Serbian version of the OUS met our sample size threshold. However, we found that it did meet sample size threshold during the Stage 2 analysis; thus, we have included it in the tables indicated by an asterisk (*).

reliability coefficients (see the supplementary materials for psychometric reports). In the main text of the manuscript, we summarize key findings across all versions. *Mplus* version 8.9 was used exclusively to conduct alignment optimization analyses, and R was used to conduct all other analyses (see the supplementary code and reports for full list of packages and versions). As such, our report meets two goals: providing a bird's-eye view of the evidence across versions and providing in-depth reporting for each instrument. This facilitates the highest possible reuse of the materials and instrument versions.

Descriptive Statistics

Descriptive Item Analysis

We computed the following statistics for each translated version and included them in that version's psychometric report:

1. Item means, medians, and variances.
2. Item response distributions (histograms, skew, and kurtosis).
3. Item correlation matrices with 95% CIs (inter-item, item-subscale, item-total, and subscale-total correlations). We indicated any correlation point estimates that are negative, weak ($r < .25$), or not statistically significant (two-tailed $\alpha = .05$) between items on the same subscale and between items and their respective subscale total.
4. Metrics for assessing multivariate normality (Q–Q plots, Shapiro–Wilk tests, Henze–Zirkler tests).

Participant Demographics

We summarized the following demographic information for each translated version and included them in that version's psychometric report:

1. Means, Medians, and Standard deviations of age.
2. Frequency count of gender.
3. Frequency count of education.
4. Frequency count of International Organization of Standardization (ISO3) country code of data collection labs' country.
5. Frequency count of religiosity.

Single-Sample Confirmatory Factor Analysis

We tested the hypothesized two-factor correlated model of utilitarianism as specified by Kahane et al. (2018) for each version's sample (see Figure 1). We also tested two, one-factor models corresponding to the Impartial Beneficence and Instrumental Harm subscales, respectively. We identified the models via variance standardization (fix the variance of each factor to 1 and estimate all loadings). All items were measured on 7-point scales, which is greater than the minimum of five recommended to acceptably treat the scale as continuous for CFAs (e.g., Rhemtulla et al., 2012). Regardless of whether the item response distributions and/or multivariate normality checks demonstrated evidence of multivariate non-normality, we estimated the models using maximum likelihood with robust standard errors and Yuan–Bentler scaling (MLR; Yuan & Bentler, 2000). In the main text, we report the fit statistics for each version in a summary table (Table 4), whereas the psychometric report for each

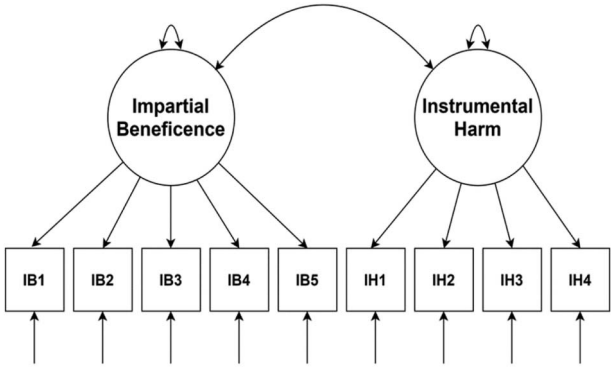

Figure 1. Two-factor OUS utilitarianism model.

Table 4. One-sample CFA model fit indices

Language	*N*	χ^2	*p*-value	CFI	TLI	RMSEA [95% CI]	SRMR
English	6,325	549.16	<.001	0.92	0.89	0.06 [0.06, 0.06]	0.05
Chinese	883	173.92	<.001	0.91	0.87	0.08 [0.07, 0.10]	0.05
French	1,096	172.43	<.001	0.89	0.85	0.08 [0.07, 0.09]	0.06
German	2,605	265.45	<.001	0.92	0.88	0.06 [0.06, 0.07]	0.04
Greek	399	36.36	.09	0.98	0.97	0.03 [0.00, 0.06]	0.04
Hungarian	777	254.04	<.001	0.77	0.68	0.11 [0.10, 0.12]	0.08
Italian	401	69.71	<.001	0.90	0.86	0.07 [0.05, 0.09]	0.05
Polish	1,024	117.93	<.001	0.93	0.90	0.06 [0.05, 0.08]	0.05
Portuguese	589	72.35	<.001	0.94	0.92	0.06 [0.04, 0.07]	0.05
Romanian	630	65.03	<.001	0.95	0.93	0.05 [0.04, 0.07]	0.04
Russian	414	52.46	<.001	0.95	0.93	0.05 [0.03, 0.07]	0.05
Serbian*	424	56.93	<.001	0.94	0.92	0.06 [0.04, 0.08]	0.06
Slovak	461	38.14	.06	0.97	0.96	0.03 [0.00, 0.06]	0.04
Spanish	869	116.00	<.001	0.91	0.88	0.07 [0.05, 0.08]	0.05
Turkish	1,111	97.14	<.001	0.94	0.92	0.05 [0.04, 0.06]	0.05

Note. Third-party reports of sample size for the Stage 1 manuscript did not indicate that the Serbian version of the OUS met our sample size threshold. However, we found that it did meet sample size threshold during the Stage 2 analysis; thus, we have included it in the tables indicated by an asterisk (*).

version includes full output from the model including parameter estimates.

Exploratory Analysis

In response to the Stage 1 review, we registered to explore model modifications on the following eight language versions with sample sizes sufficient for cross-validation (i.e., $n > 800$): US English, Chinese, French, German, Hungarian, Polish, Spanish, and Turkish (see the supplemental materials for the Stage 1 protocol; Hungarian did not meet this threshold after exclusions [$n = 777$]). Two of the leading authors reviewed the qualitative feedback from expert reviewers, modification indices, and theoretical interpretability/plausibility to identify any reasonable modifications, focusing on possible cross-loadings, correlated residuals, or item removal. For each modification, we could produce multiple plausible explanations that ranged from simple wording effects to theoretical changes to the construct. Without further qualitative and quantitative data, we could not lend any evidence to these hypotheses and thus determined that there were not theoretically sound modifications we could make that would result in increased score validity. We do not report on any modified versions in the manuscript but do include all modification indices in the version-specific reports. We also report the set of items that is invariant across the greatest number of languages as determined by alignment optimization.

Qualitative Feedback: Translator Review of Items

For each version, a bilingual expert reviewer with experience translating psychological measures read the English version of each scale and its translated version and provided qualitative feedback on the translated items via feedback sheets. We asked the reviewer to rate the quality of the translation by ranking the top 2 items for each subscale, indicating problematic items, and optionally providing qualitative feedback: "Please select the top two items for accuracy of the translation, with 1 being your top choice and 2 being your second choice. If you feel any of the items are particularly problematic in their translation, please indicate with an X and leave comments about the items, or specific words or meanings that could be confused in the translated version." Each reviewer was offered a $50 USD stipend as compensation. For each version, we report the top 2 items in each subscale (Table 5).

We also conducted a modified version of this procedure for the original English version. A coauthor who is a native English speaker reviewed each item and provided qualitative feedback on if they thought an item was problematic in any way.

Additionally, we report the problematic items indicated by reviewers for each subscale for each translated version (Table 5) and the total frequency of each item being indicated as problematic across all translated versions. After all feedback sheets were completed, we planned to review all qualitative feedback and code the main themes of the problematic items, from which we would identify categories to summarize in a table. However, there were only six items across all versions marked as problematic and full comments are included in the Results section.

Table 5. Items rated most accurately translated and problematic

Item	ZH-S	FR	DE	EL	HU	IT	PL	PT	RO	RU	SERB*	SK	ES	TR	Total ranked #1	Total ranked #2	Total flagged
Impartial beneficence (IB)																	
IB1	2	2			1		1	2	2		1		X		3	4	1
IB2	1				2					2	2	2	2		1	5	0
IB3			2	2		1	2	1		1			X		3	3	1
IB4			1	1		2						1	X	1	4	1	1
IB5		1							1				1	2	3	1	0
Instrumental harm (IH)																	
IH1	1				X						X	1	2	1	3	1	2
IH2		2	1	2	1			2	2	1	1		1		5	4	0
IH3	2	1			X	2	1			2		2	X	2	2	5	2
IH4			2	1	2	1	2	1	1		2				4	4	0

Note. The languages are abbreviated as follows: Chinese (Simplified) – ZH-S, French – FR, German – DE, Greek – EL, Hungarian – HU, Italian – IT, Polish – PL, Portuguese – PT, Romanian – RO, Russian – RU, Serbian – SERB, Slovak – SK, Spanish – ES, and Turkish – TR. Items rated as the most accurately translated are denoted by a 1. Items rated as the second most accurately translated are denoted by a 2. Items that are deemed problematic are denoted by an X. In Russian measurement invariance models, IB2 was used as an anchor item instead of IB3 as IB3 was flagged for a nonsignificant correlation with IB1. Third-party reports of sample size for the Stage 1 manuscript did not indicate that the Serbian version of the OUS met our sample size threshold. However, we found that it did meet sample size threshold during the Stage 2 analysis; thus, we have included it in the tables indicated by an asterisk (*).

Measurement Invariance Across Translations

Multiple-Group Confirmatory Factor Analysis

We evaluated the measurement invariance of each translation relative to the original English version by conducting a sequence of multiple-group CFAs with the original English version as the reference group. We selected anchor items for each subscale based on the top-rated item from the expert review of the translated version. We anticipated two potential contingencies for selecting an anchor item: first, if a reviewer could not identify a top ranked item, indicating all items as problematic, and second, if the top ranked item from a translator is flagged for poor psychometric properties during item analysis. In either case, we selected the item with the most similar loading to the original version from the single-sample CFA results. These contingencies were not realized except in Russian, and the top ranked item was flagged for nonsignificant correlation with another item on the same subscale, so the second ranked item, which was not flagged, was used as the anchor item.

For each translated version, we report the results of a complete measurement invariance analysis (using US English as the reference group), working from configural invariance of the original hypothesized model, then testing the loadings for equality (metric), followed by the intercepts (scalar), and ending with the residuals (strict). In Table 6, we report a summary table that includes model fit statistics for the final invariance model, including model χ^2 (Yuan–Bentler scaled versions), CFI (robust), TLI (robust), RMSEA with 90% CIs (robust), and SRMR. Model fit differences between measurement invariance models were also computed using model χ^2 (likelihood ratio test) and differences in CFI (robust) and RMSEA (robust). In the supplemental psychometric report, we provide full output for each invariance model, regardless of whether it was achieved. The reported level of invariance achieved using MGCFA is based on the χ^2 model fit for configural invariance using the permutation method (fail if $p < .05$). Metric, scalar, and/or strict invariance was rejected if both the χ^2 model fit difference test was rejected and the higher-order model resulted in either an RMSEA increase of .015 or greater or a CFI decrease of .01 or greater (Chen, 2008).

Configural Invariance

The configural invariance model has no constraints on parameters beyond those required for identification and setting the scale (i.e., factor means in both groups fixed to 0, loadings of anchor items in each subscale fixed to 1). If the scaled χ^2 had a p-value less than .05, we tested whether failure of configural invariance was solely due to an *overall discrepancy* (i.e., the correct specification is the same for both groups, but the model we fit was misspecified), or at least partly due to a *group-specific discrepancy* (i.e., the correct specification is different for each group, and thus, we specified the model incorrectly for at least one group) using the permutation method, which presents better Type I error rate control than conventional model fit measures (Jorgensen et al., 2018). We used 1,000 iterations.

The permutation method involves randomly assigning group membership without replacement across multiple iterations to generate empirical sampling distributions

Table 6. Multiple-group CFA measurement invariance results per language

Language	Configural invariance achieved	Final level of invariance achieved	χ^2 (final model	$\Delta\chi^2$ (final – previous)	CFI	RMSEA (90% CI)	SRMR
Chinese	No	—	727.99	—	0.92	0.06 [0.06, 0.07]	0.04
French	Yes	Metric	746.79	21.38	0.91	0.06 [0.06, 0.06]	0.04
German	No	—	822.51	—	0.92	0.06 [0.06, 0.06]	0.04
Greek	No	—	608.36	—	0.93	0.06 [0.05, 0.06]	0.04
Hungarian	No	—	808.46	—	0.90	0.07 [0.06, 0.07]	0.05
Italian	Yes	Metric	641.17	6.14	0.92	0.06 [0.05, 0.06]	0.04
Polish	No	—	661.31	—	0.92	0.06 [0.06, 0.06]	0.04
Portuguese	No	—	637.40	—	0.92	0.06 [0.06, 0.06]	0.04
Romanian	No	—	615.41	—	0.92	0.06 [0.06, 0.06]	0.04
Russian	Yes	Configural	605.80	—	0.92	0.06 [0.06, 0.06]	0.04
Serbian*	No	—	603.18	—	0.92	0.06 [0.06, 0.06]	0.04
Slovak	Yes	Metric	603.15	11.97	0.93	0.06 [0.05, 0.06]	0.04
Spanish	Yes	Metric	687.65	9.76	0.92	0.06 [0.05, 0.06]	0.04
Turkish	Yes	Metric	707.40	52.7	0.92	0.06 [0.05, 0.06]	0.04

Note. Level of invariance achieved using MGCFA was based on the χ^2 model fit for configural invariance using the permutation method (fail if $p < .05$). Metric, scalar, and/or strict invariance will be rejected if both the χ^2 model fit difference test is rejected and the higher-order model results in an RMSEA increase of .015 or greater/CFI decrease of .01 or greater (Chen, 2008). The fit statistics displayed for languages that did not meet configural invariance are from the configural invariance model. Third-party reports of sample size for the Stage 1 manuscript did not indicate that the Serbian version of the OUS met our sample size threshold. However, we found that it did meet sample size threshold during the Stage 2 analysis; thus, we have included it in the tables indicated by an asterisk (*).

of model fit measures (χ^2, CFI, RMSEA), assuming measurement invariance holds. These empirical sampling distributions can then be used to conduct statistical hypothesis tests on the model fit measures. If the null hypothesis that the correct specification is the same for both groups cannot be rejected (i.e., χ^2 $p > .05$), then the measurement invariance of the translated version was also analyzed using alignment optimization because configural invariance is a core assumption of the alignment method (in the sense that any misfit is due to an overall discrepancy from the correct specification rather than a group-specific discrepancy; see the Alignment Optimization section). If configural invariance was not achieved in the permutation method, then the translated version was excluded from further measurement invariance testing with both the multiple-group CFAs and alignment optimization.

Metric Invariance

Building on the configural invariance model, the metric invariance model constrained all factor loadings to be equal across versions. We did not test for scalar invariance or strict invariance if the likelihood ratio test comparing the metric invariance model to the configural invariance model was statistically significant and the metric invariance model resulted in either an RMSEA increase of .015 or greater or a CFI decrease of .01 or greater.

Scalar Invariance

Building on the metric invariance model, the scalar invariance model constrained all factor loadings and intercepts to be equal across versions. Additionally, the factor mean of the translated version was freely estimated. We did not test for strict invariance if the likelihood ratio test comparing the scalar invariance model to the metric invariance model was statistically significant and the scalar invariance model resulted in either an RMSEA increase of .015 or greater or a CFI decrease of .01 or greater.

Strict Invariance

The strict invariance model constrained all factor loadings, intercepts, and item error variances to be equal across versions. Additionally, the factor mean of the translated version was freely estimated. If the likelihood ratio test comparing the scalar invariance model to the strict invariance model was statistically significant and the strict invariance model resulted in either an RMSEA increase of .015 or greater or a CFI decrease of .01 or greater, we concluded that the OUS lacks strict invariance in the language of interest.

Noninvariance Effect Sizes

Regardless of the level of invariance achieved, for each two-group comparison, we computed effect sizes of noninvariance for each item using indices created by Nye and Drasgow (2011). The first index, *dMACS*, is a standardized

measure of violations of group equivalence in both factor loadings and item intercepts. This measure uses squared group differences in predicted scores given one's placement on the latent continuum to sum together all sources of bias. Consequently, noninvariance in opposite directions (either on different items or on different parts of the response distribution of a single item) will not cancel out. We follow empirically derived guidelines regarding *dMACS* estimates of .20–.39, .40–.69, and .70 or higher as representing small, medium, and large degrees of noninvariance, respectively (Nye et al., 2019). We also use ΔMean, an unstandardized measure of how much bias a given item introduces to the comparison of observed mean composite scores. Summing ΔMean across items yields the net bias in group mean comparisons and therefore accounts for reversals in the direction of noninvariance across items. The magnitude of ΔMean is interpreted in terms of the percentage of observed group mean differences that are attributable to noninvariance. The results are summarized in Table 7.

Unbalanced Sample Sizes

In their simulation study, Yoon and Lai (2018) found that unbalanced sample sizes across groups can potentially mask measurement noninvariance. To account for this, we reconducted the configural, metric, scalar, and strict invariance analyses using their recommended resampling procedure. For each language comparison, we computed 100 random samples and reported the mean $\Delta\chi^2$, ΔCFI, and ΔRMSEA values.

Alignment Optimization

In addition to using MGCFAs, we also investigated the measurement invariance of all eligible translated versions simultaneously using alignment optimization (Asparouhov & Muthén, 2014). Alignment optimization produces a factor model that is sufficient to make factor mean comparisons by selecting factor means and variances that minimize measurement noninvariance of loadings and intercepts. This can be accomplished without having to achieve scalar invariance or identify a partial measurement invariance model. It was only performed on the English version and any versions that achieved configural invariance via the permutation test.

Under alignment optimization, the configural model must be identified via variance standardization, and the alignment configuration will depend on how many versions are analyzed. There are two identification options: If only two versions were analyzed, we would fix the factor mean and variance of the English version fixed to 0 and 1, respectively. If more than two versions were analyzed, then we would freely estimate the factor mean and variance of the English version. We also use the same MLR estimation as the other factor analyses.

To evaluate the performance of the alignment procedure, we followed guidelines suggested by Asparouhov and Muthén (2014). These guidelines are based on the final step of alignment optimization, which is an item-level testing algorithm that produces significance tests and measurement noninvariance effect size estimates (R^2) for all item loadings and intercepts across groups. Performance

Table 7. dMACS estimates of noninvariance bias on Mean OUS sum scores

	Oxford Utilitarianism Scale										
	Impartial beneficence (IB)						Instrumental harm (IH)				
Language	1	2	3	4	5	Δ*M*	1	2	3	4	Δ*M*
Chinese	0.13	0.98	0.49	0.54	0.87	−5.04	0.48	0.48	0.20	0.51	−1.06
French	0.03	0.13	0.05	0.23	0.09	0.28	0.50	0.13	0.56	0.17	−2.03
German	0.09	0.19	0.21	0.45	0.09	0.52	0.12	0.14	0.25	0.18	−0.25
Greek	0.24	0.25	0.23	0.58	0.23	−0.52	0.10	0.29	0.10	0.26	−0.14
Hungarian	0.15	0.36	0.31	0.06	0.08	−0.91	0.46	0.07	0.40	0.35	0.05
Italian	0.32	0.32	0.09	0.36	0.32	1.22	0.34	0.04	0.43	0.24	−1.86
Polish	0.18	0.23	0.22	0.03	0.58	−1.64	0.34	0.12	0.23	0.44	−1.30
Portuguese	0.06	0.35	0.20	0.14	0.21	−0.18	0.41	0.43	0.53	0.15	−2.65
Romanian	0.10	0.37	0.44	0.48	0.16	−1.18	0.11	0.39	0.20	0.24	−0.22
Russian	0.60	1.01	0.55	0.47	0.76	−3.49	0.34	0.04	0.33	0.19	−0.98
Serbian*	0.08	0.32	0.37	0.35	0.07	−0.42	0.41	0.33	0.31	0.41	−1.20
Slovak	0.10	0.20	0.16	0.16	0.14	0.28	0.01	0.60	0.08	0.28	1.50
Spanish	0.18	0.16	0.28	0.16	0.07	0.78	0.19	0.11	0.30	0.19	−0.90
Turkish	0.51	0.61	0.22	0.22	0.39	1.51	0.41	0.53	0.08	0.68	−1.12

Note. Estimates are relative to the original English version (reference group). Third-party reports of sample size for the Stage 1 manuscript did not indicate that the Serbian version of the OUS met our sample size threshold. However, we found that it did meet sample size threshold during the Stage 2 analysis; thus, we have included it in the tables indicated by an asterisk (*).

is deemed adequate if no more than 25% of items are deemed noninvariant via the item-level significance tests (Muthén & Asparouhov, 2014), interpretation of the R^2 effect sizes of invariance, and interpretation of item-level noninvariance deviations. An R^2 value near 1 indicates complete invariance because the variability in item parameters is completely explained by group mean differences, whereas 0 indicates that group mean differences explain none of the variability in the item parameter. Items that appear to be highly noninvariant (i.e., significant measurement noninvariance test, *low* R^2) will be reported to help inform future partial measurement invariance analyses, if applicable. We summarize key findings from the alignment optimization here (see Table 8); full output for the alignment optimization analysis is provided in the supplementary materials.

Reliability Analysis

As per Flora (2020) and Kelley and Pornprasertmanit (2016), we computed continuous ω coefficients with bootstrapped 95% CIs (we registered 5,000 iterations but had to increase to 6,500, which is the minimum required for our sample size) for each OUS subscale, treating each subscale as a unique unidimensional measure that corresponds with the two-factor structure of the OUS proposed in Figure 1. In consideration of common conventions for reliability coefficients in psychology, we also computed Cronbach's α with bootstrapped 95% CIs (6,500 iterations) for each subscale. Estimates and confidence intervals of ω coefficients for the subscales are visualized in Figure 4a and b.

Results

The results of the single-sample CFAs are displayed followed by the results of the qualitative feedback from translators, the results from the MGCFAs, the results from the alignment optimization analysis, and the reliability analyses. The full results on each version of the OUS are included in the psychometric reports in the supplemental materials.

Single-Sample CFA

The results of the single-sample CFAs are summarized in Table 4. Overall, the single-sample CFAs revealed statistically significant evidence of misspecification. There was moderate to acceptable fit on at least one approximate fit index according to Hu and Bentler's (1999) conventional cutoffs in most languages, although their simulation conditions do not necessarily match our own and there is no straightforward relationship between degree of misfit and degree of misspecification (Greiff & Heene, 2017). The two languages in which χ^2 was not significant were Greek (p = .09) and Slovak (p = .06).

Qualitative Feedback

The results of the qualitative feedback from translators are summarized in Table 5. Throughout the translated versions, only two reviewers flagged items: Hungarian flagged IH1 and IH3 both for issues related to grammar and Spanish flagged IB1, IB3, IB4, and IH3 for issues related to grammar. Full qualitative feedback including comments can be found in the supplementary psychometric reports.

Multiple-Group Confirmatory Factor Analysis

The results of the MGCFAs are summarized in Table 6. Six of the 15 languages met configural invariance, and no languages met scalar or strict invariance.

Table 8. Alignment optimization summary table

	Oxford Utilitarianism Scale					
	Impartial beneficence (IB)			Instrumental harm (IH)		
Language	Noninvariant items	Factor mean	Factor variance	Noninvariant items	Factor mean	Factor variance
English	IB2, IB4	−1.368	1.000	IH2	−0.707	1.000
French	—	−1.230	1.100	—	−1.487	1.140
Italian	IB2, IB3	−0.533	0.883	—	−1.178	1.318
Russian	IB2	−2.957	2.536	—	−1.175	1.297
Slovak	—	−1.040	0.992	—	−0.565	0.921
Spanish	—	−1.081	1.127	IH2	−1.168	1.308
Turkish	IB1, IB2, IB3, IB4	−0.283	1.226	IH2, IH3	−1.292	1.457

Note. Noninvariant items refer to items with either noninvariant loadings or intercepts.

Noninvariance Effect Sizes

The results of the noninvariance effect sizes are summarized in Table 7. The dMACS index can be thought of as a measure of the contribution of noninvariance to expected score differences in each item, averaged across the latent distribution. It is calculated by freely estimating the item intercept and factor loading in each group, and then observing the extent to which average predicted responses differ across groups when the latent score is the same (Nye & Drasgow, 2011). Although the index is standardized, it has no absolute interpretation. The effect size conventions we use were based on simulation studies (Nye et al., 2019). The ΔMean index is intended to show what practical difference noninvariance will make in the common context where researchers are computing mean differences based on observed composites (Nye & Drasgow, 2011). In particular, ΔMean represents group mean differences in raw total scores that are attributable to noninvariance. We find that when there is noninvariance in items, it is generally of a small or medium degree, with only IB2 and IB5 having a large degree of noninvariance in the Chinese and Russian versions of the OUS. There were no striking patterns in which items consistently caused the largest amount of noninvariance across languages.

Alignment Optimization

The results of the alignment optimization analysis are summarized in Table 8. Figures 2 and 3 visualize the ranking of latent factor means for each version of the IB and IH scales, respectively.

Reliability Analysis

The results of the reliability analysis for each subscale are visualized in Figure 4. None of the metrics for reliability meet the tacit .8 cutoff with most estimates ranging between .5 and .7. Estimates and confidence intervals of Cronbach's α were not substantially different from those of the ω coefficients.

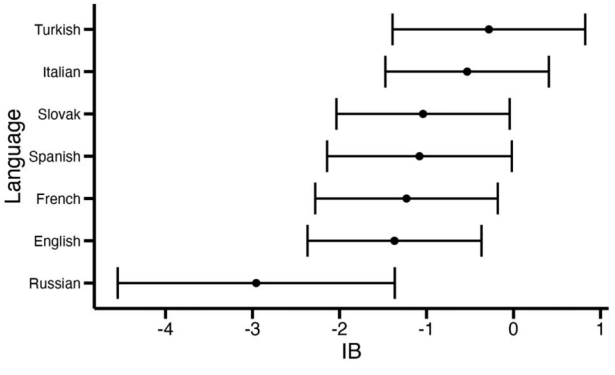

Figure 2. Factor *M* and *SD* of the impartial beneficence subscale. *Note.* The interval indicates one standard deviation above and below the factor mean for languages that met configural invariance.

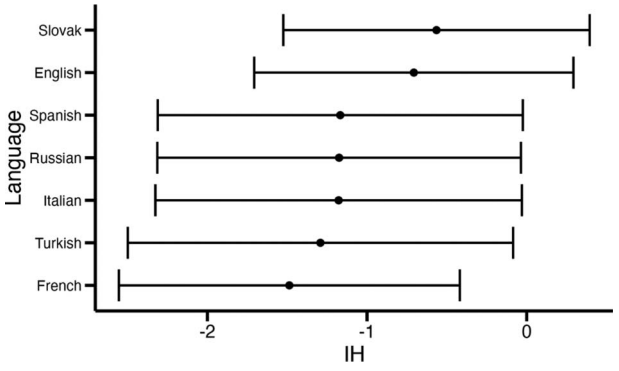

Figure 3. Factor *M* and *SD* of the instrumental harm subscale. *Note.* The interval indicates one standard deviation above and below the factor mean for languages that met configural invariance.

Discussion

We set out to assess whether the OUS can be used to measure the same underlying beliefs across numerous languages and cultural contexts in nonexpert populations and provide comprehensive psychometric reports for each version that would facilitate valid reuse of open data and materials. We found that many of the translated versions were not psychometrically equivalent. Here, we discuss implications for the construct of utilitarianism, recommendations for use of the reports and translated instruments, and future lines of research.

The Construct of Utilitarianism

The results indicate that the translated versions of this instrument do not function equivalently, and in some languages, the psychometric qualities are poor. We found that Chinese, German, Greek, Hungarian, Polish, Portuguese, Romanian, and Serbian versions did not have the same configuration of items to factors than the original English version. Overall, it appears the group-specific discrepancies range from minor issues such as differences in error covariances to major issues such as differences in the number of factors. For example, the Greek version had only three significant modification indices (one cross-loading and two error covariances), suggesting only minor issues with the item configuration. On the other hand, the Chinese, Polish, and German versions each had 21 modifications

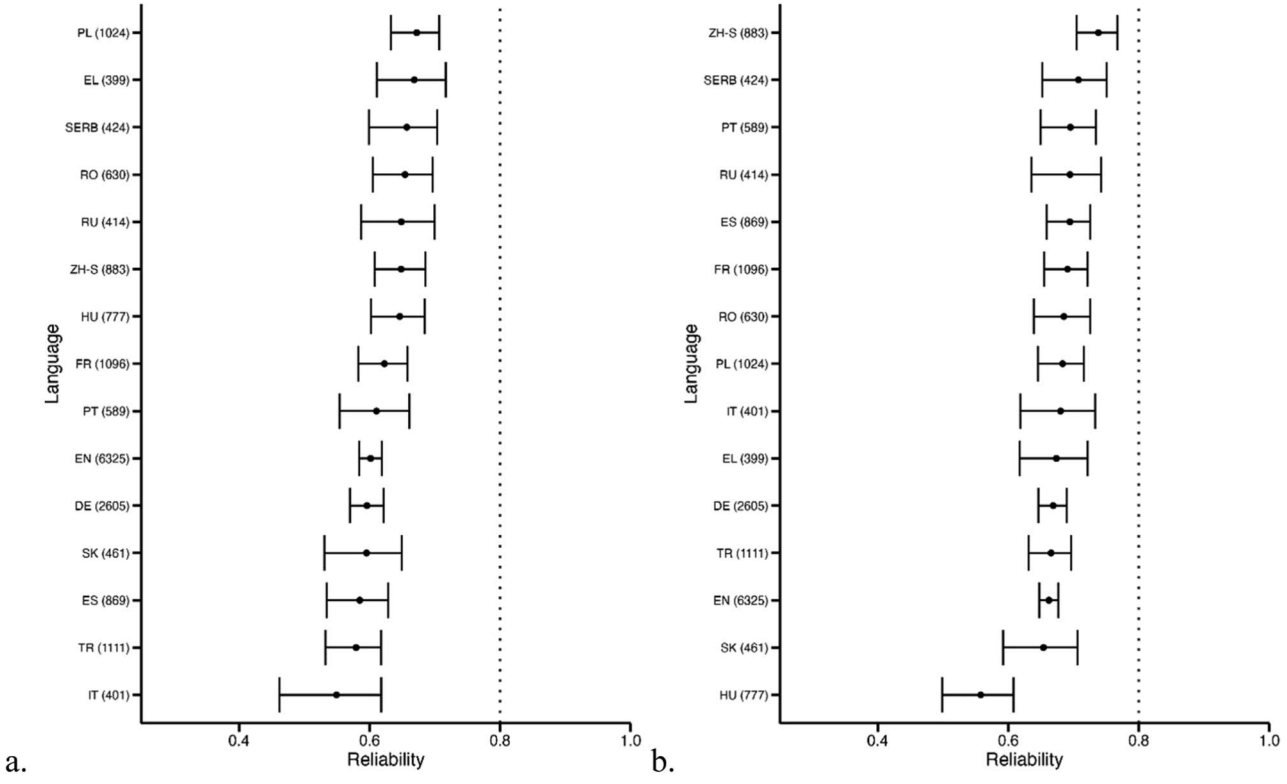

Figure 4. Reliability estimates of the subscales. *Note.* a depicts the reliability estimates of the Impartial Beneficence subscales. b depicts the reliability estimates of the Instrumental Harm subscales. Both figures depict the ω estimate and 95% confidence intervals. The dotted line indicates the tacit .8 cutoff used for reliability. The estimate and 95% confidence interval for Cronbach's α is not substantially different from that depicted here.

with significant modification indices, suggesting major misfit of the fitted item configuration. For these languages with many significant modification indices, we would suggest further investigation of the underlying model through analyses such as exploratory factor analysis. The differences in group-specific discrepancies suggest that the sources of misspecification could be idiosyncratic to each version, representing wording effects or evidence of cultural differences in the construct itself. Despite that we report a large-scale analysis from a rich data source, this study cannot address the qualitative differences between versions or the underlying cultural differences that may contribute to our results. For example, when we reviewed these results, we would generate multiple theories as to why some languages showed better fit than others. However, in the absence of cultural expertise, qualitative inquiry, and a follow-up study, we did not consider any of our theories particularly compelling. An important next step for this psychometric research is research into the construct itself, how it differs across cultures, and how people speaking different languages may or may not interpret the items differently (e.g., Renberg et al., 2008).

Recommendations for Use

Of the instruments we tested, six demonstrated configural equivalence to the original version. For the rest of the instruments that did not, this means that the configuration of items to factors differs between the original and translated versions. This could manifest as different sets of items forming different factors, different patterns of error covariance, or unmodeled cross-loadings. The fact that six of the languages had no group-specific discrepancies suggests that Impartial Beneficence and Instrumental Harm are constructs that exist and have similar manifestations across many cultures.

For the versions that did not meet configural invariance, there is no evidence their observed scores can be used to make valid comparisons across groups. We do not recommend that the Chinese, German, Greek, Hungarian, Polish, Portuguese, Romanian, or Serbian versions be used to make comparisons to the original version. These groups may be comparable in a latent variable framework, as some small amount of configural nonequivalence could be modeled. However, this would require follow-up analysis to determine the source of the misfit and if a viable model can be found. Despite that these versions were not

equivalent to the English version, they could still be used for research on those groups. We recommend researchers interested in using any version to inspect the validity report for that version and consider if the psychometric evidence supports their intended use. Many of the versions have mediocre to acceptable model fit and mediocre reliability. However, it is notable that the mediocre reliability could be a result of the shortness of the scale and, if only using the OUS for group statistics, reliability at the level of the typical .8 cutoff may be unnecessary (Ziegler et al., 2014). While these properties are not optimal, they may be acceptable for use with caution or to further develop the construct and instrument.

Of those that demonstrated configural invariance, the alignment flagged at least one noninvariant item between the US English reference group and the mean of the item parameters across the six language versions. The French, Slovak, and Spanish versions demonstrated full invariance of all items with each other, which suggests that these versions can be used to make valid comparisons of the observed scores. However, for the others, we recommend a latent variable model that can account for item property differences if the research goal is to make comparisons across groups. In addition to the alignment method, we tested groups that demonstrated configural invariance for metric, scalar, and strict. Russian met only configural invariance while French, Italian, Slovak, Spanish, and Turkish met metric invariance with English. Because of this, we recommend that researchers interested in using one of these translated instruments to review the reports in depth and consider the magnitude and pattern of the results, as well as the effect size measures. For example, while the results of the MGCFA suggest varying degrees of noninvariance, the effect size reveals small item differences. These are the types of considerations we recommend researchers make when considering using these translated instruments as small amount of noninvariance may not have a practical difference on results. Researchers can use the validity reports, sensitivity analysis, and caution to use these instruments in downstream research projects.

Limitations

We highlight three serious limitations to this work. First, many of the baseline measurement models exhibited model misspecification. These models form the foundation of the latter equivalence testing and even minor amounts of misfit can increase Type I errors for detecting nonequivalence (French & Finch, 2011). Therefore, some findings of invariance may be due in part to model misspecification rather than actual, meaningful measurement invariance. The misfit is consistent with both (a) the misspecifications being small and nonsystematic and (b) the misspecifications being large enough to threaten the validity of the subscale scores. We investigated the modification indices and hypothesized sources of the misfit, but the data at hand could provide no evidence for or against our hypotheses. For example, item error covariances suggested alternative interpretations of items, but without cognitive interviewing or more research, we cannot confirm those interpretations. While these data have been useful for evaluating the current scale, there are extremely limited in their ability to contribute to improving the scale. Finally, we only consider the language the instruments were translated into, which ignores the heterogeneity within a version of the scale. For example, the Spanish version was administered in Argentinian, Chilean, Colombian, Ecuadorian, Mexican, Peruvian, Salvadorian, and European Spanish versions, which are from culturally diverse groups. It is likely that within these language groups there are cultural, demographic, and sociopolitical subpopulations that could exhibit further measurement differences. We do not recommend researchers solely consider language groups when exploring measurement differences, despite that those further analyses were not feasible for this project. Relatedly, we did not have access to representative samples, which would provide a means to understanding cultural differences in utilitarianism. The data were collected by member labs of the PSA, which are unlikely to be representative of the demographics for a given country or culture.

Future Research

Some of the limitations could be addressed with future research. The version-specific reports identify gaps in the validity evidence, which can form the basis of further development and work for the instrument. While the instrument was not developed for cross-cultural measurement at the outset, this analysis provides lot of rich detail that could launch a global validation study. Any new or revised version would benefit from new equivalence analyses and could be sample sized planned to explore further subgroups and cultures within languages. While we did not compare versions to one another in all possible combinations, focusing on comparisons to the original version, an expanded large-scale data collection would enable this. Furthermore, our results suggest there are differences between the mean level of the factors on these constructs. This could launch further research on the psychological aspects that contribute to these differences.

Although this project demonstrates the knowledge that can come out of doing validation on scales used in Big Team Science, it also demonstrates that measuring constructs in

many languages is difficult. The lack of psychometric evidence for translated scales directly negates interpretations of the data and complicates reuse. We hope that this project will motivate big teams to consider measurement early, undertake validation as a central part of the research, and ultimately turn Big Team Science into an engine for global, reusable, psychological instruments.

Conclusion

We sought to leverage Big Team Science efforts to disseminate large-scale validity evidence for translated versions of the Oxford Utilitarianism Scale. This project reveals the complex research needed to develop and use scales in many languages. Big Team Science offers many avenues for innovation, and this project reveals that global measurement of psychological constructs is a key but an underexplored one. Our hope is that this project will shift practice to expand the consideration of measurement in large-scale projects and spur the development of methods and practices for large-scale measurement.

References

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, *52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J. F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337. https://doi.org/10.1073/pnas.1911517117

Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albalooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., … Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*(6), 880–895. https://doi.org/10.1038/s41562-022-01319-5

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Erzi, S. (2019). Psychometric properties of adaptation of the Oxford Utilitarianism Scale to Turkish. *HUMANITAS - Uluslararası Sosyal Bilimler Dergisi*, *7*(13), 132–147. https://doi.org/10.20304/humanitas.507126

Everett, J. A. C., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences*, *24*(2), 124–134. https://doi.org/10.1016/j.tics.2019.11.012

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, *3*(4), 484–501. https://doi.org/10.1177/2515245920951747

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3

French, B. F., & Finch, W. H. (2011). Model misspecification and invariance testing using confirmatory factor analytic procedures. *The Journal of Experimental Education*, *79*(4), 404–428. https://doi.org/10.1080/00220973.2010.517811

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural differences in careless responding. *Journal of Cross-Cultural Psychology*, *50*(3), 336–357. https://doi.org/10.1177/0022022119827379

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in Cognitive Sciences*, *6*(12), 517–523. https://doi.org/10.1016/S1364-6613(02)02011-9

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.3389/fpsyg.2014.00980

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, 980. https://doi.org/10.3389/fpsyg.2014.00980

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/s0140525x0999152x

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jorgensen, T. D., Kite, B. A., Chen, P.-Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods*, *23*(4), 708–728. https://doi.org/10.1037/met0000152

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. https://doi.org/10.1037/a0040086

Koziol, N. A., & Bovaird, J. A. (2018). The impact of model parameterization and estimation methods on tests of measurement invariance with ordered polytomous data. *Educational and Psychological Measurement*, *78*(2), 272–296. https://doi.org/10.1177/0013164416683754

Law, K. F., Campbell, D., & Gaesser, B. (2022). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychology Bulletin*, *48*(3), 426–444. https://doi.org/10.1177/01461672211002773

Luong, R., Flake, J. K., McAuliffe, W., & Oshiro, B. (2023). *Structural validity evidence for the Oxford Utilitarianism Scale across 15 languages* [Open Data]. https://osf.io/y96rm/

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance.

Structural Equation Modeling, *14*(4), 611–635. https://doi.org/10.1080/10705510701575461

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. https://doi.org/10.1016/j.tics.2006.12.007

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology*, *5*, Article 78. https://doi.org/10.3389/fpsyg.2014.00978

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678–709. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Paris, B., IJzerman, H., & Forscher, P. S. (2020). *PSA 2020–2021 study capacity report*. https://osf.io/k6hzw/#!

Psychological Science Accelerator. (2018, February 7). *Translation process*. https://psysciacc.org/translation-process/

Renberg, T., Kettis Lindblad, Å., & Tully, M. P. (2008). Testing the validity of a translated pharmaceutical therapy-related quality of life instrument, using qualitative "think aloud" methodology. *Journal of Clinical Pharmacy and Therapeutics*, *33*(3), 279–287. https://doi.org/10.1111/j.1365-2710.2008.00921.x

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184. https://doi.org/10.1023/A:1019923923537

Slaney, K. L., & Maraun, M. D. (2008). A proposed framework for conducting data-based test analysis. *Psychological Methods*, *13*(4), 376–390. https://doi.org/10.1037/a0014269

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and against (Issue 3)*. Cambridge University Press.

Tetlock, P. E. (2003). Thinking the unthinkable: Sacred values and taboo cognitions. *Trends in Cognitive Sciences*, *7*(7), 320–324. https://doi.org/10.1016/s1364-6613(03)00135-9

Todd, B., & MacAskill, W. (2017). *Is it ever okay to take a harmful job in order to do more good? An in-depth analysis*. https://80000hours.org/articles/harmful-career

Yoon, M., & Lai, M. H. (2018). Testing factorial invariance with unbalanced samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 201–213. https://doi.org/10.1080/10705511.2017.1387859

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – five misunderstandings and ways to overcome them [Editorial]. *Journal of Individual Differences*, *35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148

History
Received November 10, 2021
Revision received October 22, 2023
Accepted November 3, 2023
Published online April 30, 2024

Section: Methodological Topics in Assessment

Acknowledgments
We thank Marijke C. Leliveld for their contributions in the Stage 1 submission: Investigation, Methodology, Resources, Writing – Review & Editing; Matej Hruška for their contributions in the Stage 1 submission: Resources, Writing – Review & Editing; and Hu Chuan-Peng, Petros Roussosm Bence Bakos, Alberto Mirisola, Gabriela Marcu, and Ilya Zakharov for their contributions in providing translation feedback.

Conflict of Interest
The authors have no conflicts of interest to disclose.

Authorship
Briana Oshiro: Formal analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review & Editing; William H. B. McAuliffe: Project administration, Methodology, Formal analysis, Writing – Original Draft, Writing – Review & Editing; Raymond Luong: Conceptualization, Formal analysis, Methodology, Writing – Original Draft, Writing – Review & Editing; Jessica Kay Flake: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Resources, Writing – Original Draft, Writing – Review & Editing; Anabela C. Santos: Investigation, Writing – Review & Editing; Andrej Findor: Investigation, Writing – Review & Editing; Anna O. Kuzminska: Investigation, Writing – Review & Editing; Anthony Lantian: Investigation, Writing – Review & Editing; Asil Özdoğru: Investigation, Writing – Review & Editing; Balazs Aczel: Writing – Review & Editing; Bojana M. Dinić: Investigation, Writing – Review & Editing; Christopher R. Chartier: Supervision, Project administration, Writing – Review & Editing; Jasper Hidding: Investigation, Writing – Review & Editing; Job A. M. de Grefte: Writing – Review & Editing; John Protzko: Project administration, Investigation, Writing – Review & Editing; Mairead Shaw: Software, Writing – Review & Editing; Maximilian A. Primbs: Investigation, Writing – Review & Editing; Nicholas A. Coles: Supervision, Writing – Review & Editing; Patricia Arriaga: Investigation, Writing – Review & Editing; Patrick S. Forscher: Writing – Review & Editing; Savannah C. Lewis: Project administration, Writing – Review & Editing; Tamás Nagy: Investigation, Writing – Review & Editing; Wieteke C. de Vries: Investigation, Writing – Review & Editing; William Jimenez-Leal: Investigation, Writing – Review & Editing; Yansong Li: Investigation, Writing – Review & Editing.

Open Science
All materials, analysis code, and psychometric reports are openly available on the Open Science Framework at https://osf.io/y96rm/ (Luong et al., 2023).

Open Data: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant.

Open Materials: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology.

Funding
Jessica Kay Flake and Raymond Luong's work was funded by the Ware's Prospector's Innovation Fund awarded to Jessica Kay Flake (Grant No. 249040, 2019). Raymond Luong's work was also

supported in part by Mitacs through the Mitacs Research Training Award, by funding from the Social Sciences and Humanities Research Council (SSHRC 430-2021-00239), and by funding from the Fonds de recherche du Québec – Société et culture.

ORCID
Briana Oshiro
https://orcid.org/0000-0002-4235-2878
Raymond Luong
https://orcid.org/0000-0001-6587-6159
Anabela C. Santos
https://orcid.org/0000-0001-7963-8397
Andrej Findor
https://orcid.org/0000-0002-5896-6989
Anna O. Kuzminska
https://orcid.org/0000-0002-6060-4549
Anthony Lantian
https://orcid.org/0000-0001-7855-3914
Asil A. Özdoğru
https://orcid.org/0000-0002-4273-9394
Bojana M. Dinić
https://orcid.org/0000-0002-5492-2188
Nicholas A. Coles
https://orcid.org/0000-0001-8583-5610
Patricia Arriaga
https://orcid.org/0000-0001-5766-0489
Savannah C. Lewis
https://orcid.org/0000-0002-9948-1195
Tamás Nagy
https://orcid.org/0000-0001-5244-0356
Wieteke C. de Vries
https://orcid.org/0000-0003-1511-0424
William Jimenez-Leal
https://orcid.org/0000-0002-8824-5269
Jessica Kay Flake
https://orcid.org/0000-0002-3498-615X

Briana Oshiro
Department of Psychology
McGill University
2001 McGill College Avenue
Montreal
QC H3A 1G1
Canada
briana.oshiro@mcgill.ca