Journal of Religion and Health
https://doi.org/10.1007/s10943-021-01351-1

ORIGINAL PAPER

The Negative Religiousness-IQ Nexus is a Jensen Effect on Individual-Level Data: A Refutation of Dutton et al.'s 'The Myth of the Stupid Believer'

Edward Dutton[1] · Emil Kirkegaard[2]

Accepted: 15 July 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

A recent study by Dutton et al. (J Relig Health 59:1567–1579. https://doi.org/10.1007/s10943-019-00926-3, 2020) found that the religiousness-IQ nexus is not on *g* when comparing different groups with various degrees of religiosity and the non-religious. It suggested, accordingly, that the nexus related to the relationship between specialized analytic abilities on the IQ test and autism traits, with the latter predicting atheism. The study was limited by the fact that it was on group-level data, it used only one measure of religiosity that measure may have been confounded by the social element to church membership and it involved relatively few items via which a Jensen effect could be calculated. Here, we test whether the religiousness-IQ nexus is on *g* with individual-level data using archival data from the Vietnam Experience Study, in which 4462 US veterans were subjected to detailed psychological tests. We used multiple measures of religiosity—which we factor-analysed to a religion-factor—and a large number of items. We found, contrary to the findings of Dutton et al. (2020), that the IQ differences with regard to whether or not subjects believed in God are indeed a Jensen effect. We also uncovered a number of anomalies, which we explore.

Keywords Religion · Intelligence · Cognitive ability · Jensen effect · Differential item functioning · Local structural equation models · Item response theory

✉ Edward Dutton
e.c.dutton@dunelm.org.uk

Emil Kirkegaard
emil@emilkirkegaard.dk

[1] Asbiro University, Lodz, Poland

[2] Ulster Institute for Social Research, London, UK

Published online: 26 July 2021

Introduction

Many studies have found a weak negative relationship between religiousness and IQ. The first studies reporting this finding were published in the 1920s (e.g. Gilkey, 1924; Howells, 1928), and it has been replicated ever since. Meta-analyses have shown that this relationship is in the region of − 0.2, in the general population, when using 'religious belief' as a measure, and − 0.1 when employing 'religious attendance' (e.g. Zuckerman et al., 2013). A recent meta-analysis has, once more, found that the relationship between religious belief and IQ is approximately − 0.2 (Zuckerman et al., 2020), and a meta-analysis of measures of reflective thinking similarly found a negative association of − 0.18 (Pennycook et al., 2016). This relationship has also been replicated using the very large OKCupid dataset with 33–37 k subjects in the main regressions (Kirkegaard & Lasker, 2020) using a religiousness factor based on five questions. The standardized beta in the final model (controlling for age, sex, race, sexual orientation, and country/state) was −0.24. Similar weak negative correlations are also found between many measures of religiousness and assorted proxies for intelligence, such as education level and salary (Meisenberg et al., 2012). The religious groups that are more fundamentalist (more fervent and dogmatic in their religious beliefs) tend to have lower average IQ than do groups that are more religiously liberal (Nyborg, 2009).

A variety of theories have been developed to explain this consistent relationship such as: (1) everybody needs the certainty of a consistent worldview and if people are insufficiently intelligent to follow a purely scientific one then they will retreat into religion (Nyborg, 2009). (2) The arguments for God's existence are illogical, meaning that intelligent people would be better able to see through them (Dutton, 2014). (3) We are adapted to the Savanna, which is 'evolutionarily familiar', where we solved problems using instinct and there we developed religious belief or, at least, belief in a spiritual universe. Moving off the Savanna, we could no longer solve problems using instinct, so had to use intelligence. Thus, intelligent people are attracted to other 'evolutionarily novel' ways of thinking, such as atheism (Kanazawa, 2012). (4). A component of problem-solving, and thus of intelligence, involves the ability to rise above our instincts, no matter which ecology they have derived from, and test out non-instinctive, superficially odd possibilities in pursuit of solving a problem. Intelligent people will, therefore, be attracted to multiple unusual ways of thinking, including atheism (Dutton & Van der Linden, 2017). Proponents of these models reject the idea that secular ideologies are more logical than belief in God, arguing that both involve non-empirical dogmas and an implicit belief in fate, and also by cautiously defending versions of William James' 'pragmatic argument' for believing in God (Dutton & Van der Linden, 2017). But the problem with each of these explanations is that they assume that the nexus really does relate to intelligence; that it is on the highly heritable and core intelligence ability known as *g* (general intelligence; Jensen, 1998), and not just on specialized skills. However, a recent study has provided cautious evidence that the nexus is *not* on *g*. The relationship is *not* a so-called Jensen effect.

Dutton et al. (2020) have analyzed two large data sets from the Netherlands, allowing them to compare the IQs of groups with different levels of religiousness, including those who were atheists and agnostics. They found that the religiousness-IQ nexus was not on *g,* meaning that it related to specialized abilities rather than to general intelligence. This study can be argued to have provided evidence that the relationship is not on *g,* at least when comparing religious and non-religious samples from the same ethnic group within a particular country. Evidently, the study's main limitation is its use of group-level data rather than individual data. It is also potentially limited by the nature of the data, which involved IQ tests being administered to groups of church members, agnostics, and atheists. This is because there is a social element to church membership and attendance, with intelligence predicting general engagement with civic activities (Rindermann et al., 2012). In addition, it is limited by the fact that it uses only one method of ascertaining religious belief (church membership or otherwise), it does not give people the chance to indicate the extent of their religious belief, and, moreover, even if it had done, belief in God is only one aspect of religious belief; with religions tending to involve a much more complex theology. A second problem with the Dutton et al. study is that the power is likely to be very low. The groups compared typically had sample sizes in the 100 s, and the gaps were very small, often about 2 IQ points.

A third problem was unearthed when we carried out a simulation study to estimate the statistical precision of the Dutton et al. study. We did this by simulating data from the hypothesized true case where *g* gaps entirely account for group gaps (i.e. the true Jensen correlation is 1.0). We furthermore varied sample size and the gap size. To maximize the comparability with the prior study, we used the same *g*-loadings as reported in their tables, 10 and 9 *g*-loadings, respectively, for the GIT and EMS test batteries. For each simulation setting, we simulated 100 results. Figure 1a, b shows the results.

It can be seen that at small sample sizes and at small gap sizes, the distribution of correlations seen is not close to their true value of 1.0. Thus, the method has a downwards bias as a function of gap size and sample size. This means that any study based on small gaps or with small sample sizes or both is likely to produce at best uninformative or at worst misleadingly negatively biased results. Most of the results in Dutton et al.'s analysis come from a single dataset with six group comparisons. The sample sizes in Dutton et al.'s two samples varied across the comparisons, but they ranged from a total of 190 to 544, and with gaps of about 0.7 to 5.9 IQ, thus at best about $d = 0.40$ (non-members vs. Roman Catholics, gap = 5.85 IQ, combined $n = 378$, $r = -0.49$). The average was about 320 people, thus at best 160 people per group, and a 3 IQ gap ($d = 0.2$). As a consequence of these values, the statistical precision of Dutton et al.'s results is likely to be very low, and the results thus misleading. The results in their final and large comparison (almost 9 k cases, thus at best about 4.5 k per group) were based on a group difference of about 2 IQ points, again, too small for useful precision. Our results are in line with prior simulation analyses of this method, showing that it has some known biases (Sorjonen et al., 2017). Thus,

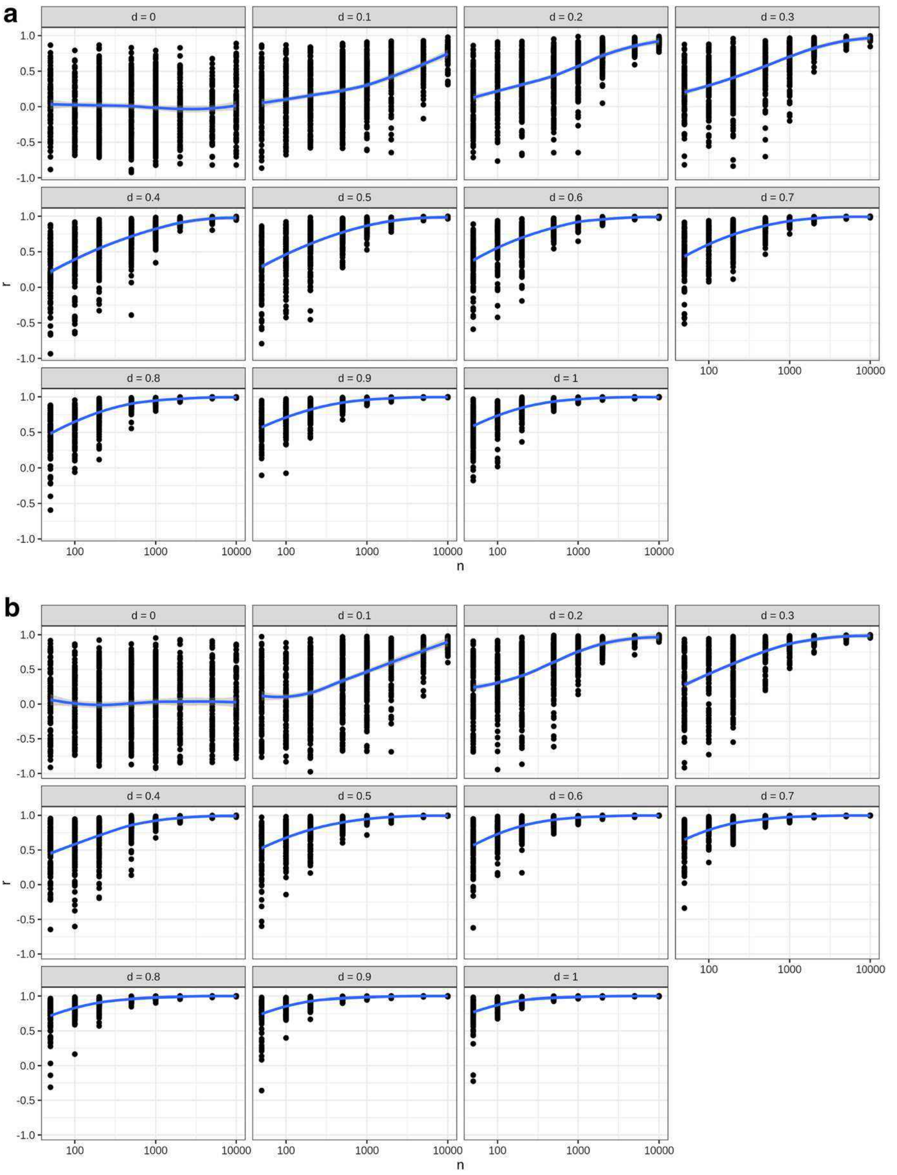

Fig. 1 **a**–**b** Simulation results for Jensen's method. **d** refers to the gap size between two equal-sized groups. The upper plot is based on GIT *g*-loadings, and the bottom plot is based on EMS loadings. Values copied from Dutton et al. Blue line is the LOESS fit. *LOESS* locally estimated scatterplot smoothing, a nonlinear smoothing function. (Color figure online)

their method suffers from a serious problem, casting doubt on whether their conclusions are accurate.

If Dutton et al.'s findings could be replicated using individual-level data and with a less problematic method, then this could be said to relatively conclusively prove that the negative religiousness-IQ nexus is not on *g*. This would be further strengthened if this was found on different aspects of religiosity—such as practice as against belief—as well as on different elements of religious belief. Accordingly, in this study, we set out to replicate the findings of Dutton et al. using individual-level data, as well as using multiple means of measuring religiosity.

Method

We used archival data from the Vietnam Experience Study (VES, https://www.cdc.gov/nceh/veterans/default1c.htm). This is a large longitudinal study of 4,462 US veterans (3,654 whites, 200 Hispanics, 525 blacks, 49 Native Americans, and 34 Asians). In terms of race, the sample was, therefore, roughly representative of the US population at the time: 82% white, 12% black, 4% Hispanic, 1 Native American, and 1% Asian. VES is a longitudinal dataset of male US military personnel who were inducted in the period 1965 to 1971. There was an extensive follow-up in 1985–1986, approximately 18 years after initial contact. The dataset was made as a case–control study of the effects of Agent Orange (chemical warfare) in the Vietnam War. Consequently, about 65% of the sample are Vietnam veterans and the rest were stationed elsewhere (e.g. Germany or Japan). Details of the recruitment and findings can be found in (The Centers for Disease Control Vietnam Experience Study, 1988a, 1988b, 1988c). The dataset includes data from 19 different cognitive tests. These have been described in detail in several previous papers and include measures of verbal reasoning, arithmetic, spatial ability, psychomotor ability, and memory (Kirkegaard & Nyborg, 2020; Nyborg & Jensen, 2000, 2001). At the follow-up, the mean age was 38 (SD 2.5).

Cognitive abilities were measured both in the initial wave and the follow-up wave. Two of the tests were given twice. The tests were as follows:

1. Grooved Pegboard Test (GPT, right hand): A measure of manual dexterity and fine motor speed (Ruff & Parker, 1993). The speed score is the reciprocal of the number of seconds taken to place a set of pegs in a grooved hole as quickly as possible.
2. GPT (left hand).
3. Paced Auditory Serial Addition Test (PASAT): A measure of mental control, speed, and computational and attentional abilities (Tombaugh, 2006). The subject mentally adds a sequence of numbers in rapid succession. Score is the total number of correct responses.
4. Rey-Osterrieth Complex Figure Drawing (CFD): A measure of visio-spatial ability and memory (Shin et al., 2006). The direct copy score (CFDD) is given from a subject reproducing a complex spatial figure, while the figure is in full view.
5. CFD, copy from immediate recall. The immediate recall score (CFDI) is given from a subject reproducing a complex spatial figure immediately after being shown it.

6. CFD, copy from delayed recall. The delayed recall score (CFDL) is given from a subject being exposed to a complex spatial figure and, after 20 min of other activities, drawing it.
7. Wechsler Adult Intelligence Scale-Revised (WAIS-R), general information (Leckliter et al., 1986). A test of general knowledge.
8. WAIS-R, block design. A test of spatial ability.
9. Word List Generation Test (WLGT). A measure of verbal fluency. The subject generates as many words as possible which begin with the letters F, A, and S for 60 s. The score is the total number of words generated.
10. Wisconsin Card Sort Test (WCST). A measure of executive function (Greve et al., 2005). The score is the ratio of correct responses to countable responses.
11. Wide Range Achievement Test (WRAT). Measures ability to read aloud a list of single words (untimed) (Witt, 1986).
12. California Verbal Learning Test (CVLT). A measure of verbal learning and memory (Elwood, 1995). The subject recalls a list of 16 words over five repeated learning trials. The score is the total correct over five trials.
13. Army Classification Battery (ACB). A verbal test administered at induction (ACBVE) (Bayroff & Fuchs, 1970).
14. ACB verbal. Administered at the follow-up interview (ACBVL).
15. ACB arithmetic reasoning test. An arithmetic test administered at induction (ACBAE).
16. ACB arithmetic. Administered at the follow-up interview (ACBAL).
17. Pattern Analysis Test (PAT). A measure of pattern recognition administered at induction.
18. General Information Test (GIT). A test of general knowledge administered at induction.
19. Armed Forces Qualification Test (AFQT). A general aptitude battery. This measure is the total score on four subtests (word knowledge, paragraph comprehension, arithmetic reasoning, and mathematics knowledge) administered at induction.

Five of the tests (13, 15, 17–19) were given at induction and the remaining at the follow-up interview.

In the second wave, the Minnesota Multiphasic Personality Inventory (MMPI) was administered (1975 version; Dahlstrom et al., 1975). This is a questionnaire of 566 statements that individuals marked as either true or false. This battery was designed to measure various aspects of mental health as well as some other traits (e.g. masculinity–femininity). We searched the list of statements for items related to religiousness and found 12 items, shown in Table 1.

We then split the questions into two categories: those six that were purely related to beliefs (58, 115, 249, 258, 373, and 483), and the rest. The remaining set of questions contain behavioural (e.g. frequency of reading the Bible), personality (e.g. lack of patience for unbelievers), or socially comparative elements (e.g. more religious than others).

Table 1 Items used to measure religiousness

Question	Item	Prevalence	Missing
53	A minister can cure disease by praying and putting his hand on your head	0.08	0.01
58	Everything is turning out just like the prophets of the Bible said it would	0.5	0.04
95	I go to church almost every week	0.25	0.00
115	I believe in a life hereafter	0.76	0.01
206	I am very religious (more than most people)	0.17	0.00
249	I believe there is a Devil and a Hell in afterlife	0.63	0.01
258	I believe there is a God	0.92	0.01
373	I feel sure that there is only one true religion	0.37	0.01
476	I am a special agent of God	0.09	0.00
483	Christ performed miracles such as changing water into wine	0.77	0.03
490	I read in the Bible several times a week	0.13	0.00
491	I have no patience with people who believe there is only one true religion	0.27	0.01

Results

We scored intelligence using exploratory factor analysis of the 19 tests, as done in prior studies using the same dataset (Kirkegaard & Nyborg, 2020; Nyborg & Jensen, 2000, 2001). Before analysis, we imputed the missing data using the IRMI algorithm in the **vim** package (Templ et al., 2015). The *g* factor accounted for 42% of the variance (minimum residuals method, scored by the regression method, using the psych package for R (Revelle, 2020)). Factor loadings are given in Table 2 (further down). Figure 2 shows the distribution of scores, which was roughly normal.

The religious data were based on dichotomous (binary) indicators, thus necessitating a more complex method. We used item response theory-based factor analysis, as implemented in the **mirt** package for R (Chalmers et al., 2020). We scored two versions of this, one with all the 12 items, and one based only on the six items concerned with beliefs (pure set). Both analyses showed a strong general factor based on a positive manifold (full set 56% variance, and 66% for the pure set). The empirical internal reliabilities were estimated at 0.84 and 0.68 (based on **mirt**'s *empirical_rxx()* function). Figure 3 shows the distributions. For the full set, the scores were roughly normal, except there was a bump on the left tail for the non-religious. The pure set was less normal.

Religiousness was negatively correlated with intelligence: -0.18 for the total score and -0.21 for the pure scale (both $p<0.001$, $\mathrm{SE}=0.0153$). Figure 4a, b shows the relationships.

The relationship was mostly linear but with evidence of diminishing returns on the left tail for the pure scale ($p=0.0006$, likelihood ratio test when compared to a natural spline model). If we take this finding seriously, it seems that IQs that are

Table 2 Test-level relationships to religiousness and factor loadings

Test	*g*-loading	*r* religiousness	*r* religiousness pure
VE time1	0.82	−0.20	−0.23
AR time1	0.81	−0.14	−0.18
PA	0.71	−0.14	−0.17
GIT	0.69	−0.19	−0.20
AFQT	0.85	−0.18	−0.21
VE time2	0.82	−0.17	−0.20
AR time2	0.82	−0.13	−0.17
WAIS BD	0.67	−0.14	−0.17
WAIS GI	0.76	−0.15	−0.20
WRAT	0.73	−0.16	−0.21
PASAT	0.57	−0.09	−0.11
WLGT	0.49	−0.09	−0.12
Copy direct	0.47	−0.06	−0.08
Copy immediate	0.55	−0.05	−0.08
Copy delayed	0.55	−0.05	−0.08
CVLT	0.42	−0.04	−0.06
WCST	0.46	−0.08	−0.09
GPT left	0.34	−0.06	−0.08
GPT right	0.33	−0.06	−0.08

over 100 are more strongly negatively related to religiousness than these below 100, and IQs below 80 are less positively related to religiousness. This kind of finding has a parallel using national intelligence data (Lynn et al., 2009). With regard to the *g*-loading of the relationship, we used Jensen's method, as done in the prior meta-analysis. Figure 5a, b shows the results visually, while Table 2 gives the numbers.

In contrast to Dutton et al., we find very strong relationships between the *g*-loading of an item and its negative correlation with religiousness, thus providing sound evidence that the relationship is mainly or entirely due to the *g* factor and not due to other cognitive abilities: $r = -0.86$ using the religiousness score from the full set, and $r = -0.91$ using the pure set scores. Aside from the *g*-loadedness of the pattern, one may wonder whether the pattern was due to plausible demographic confounders. To investigate this, we ran a series of regression models with controls added. Results are shown in Tables 3 and 4.

The person-level regression results show that the relationship is not plausible due to the confounders considered, including age, race, income, or education. The weakest slope is seen with the full religiousness scale, where education and income are controlled, but even here, the slope is −0.14. Since adding these controls is likely adjusting for a mediator, this small decrease should not be interpreted strongly. Furthermore, the standard errors are too large to have certainty that even this small decrease in the beta is a real change. The models that add interactions between race and intelligence find suggestive evidence that the relationship is weaker among blacks for the pure scale [interaction $p = 0.008$, implied slope among blacks $= -0.08$ $(-0.22 + 0.14)$]. Figure 6a, b shows the marginal effects. It should be noted here that

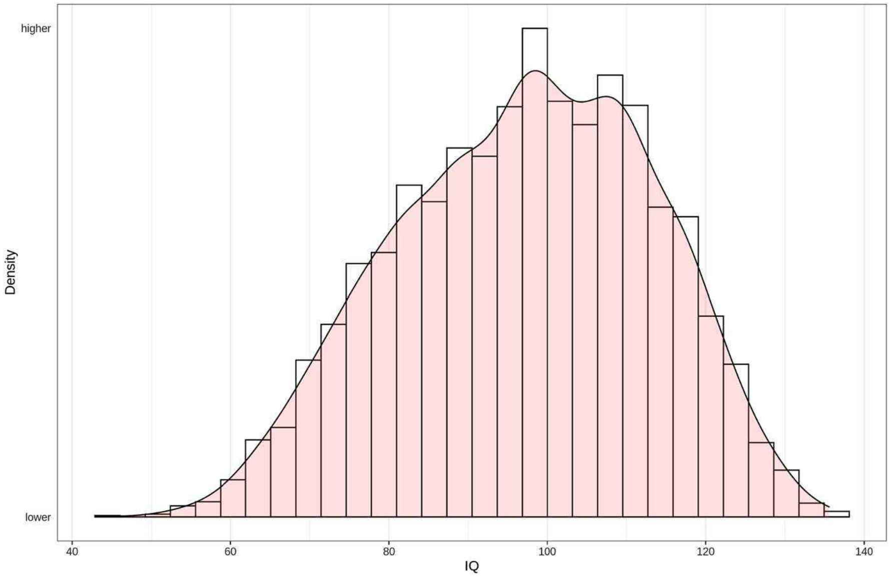

Fig. 2 Distribution of IQ scores in the dataset (white = 100/15)

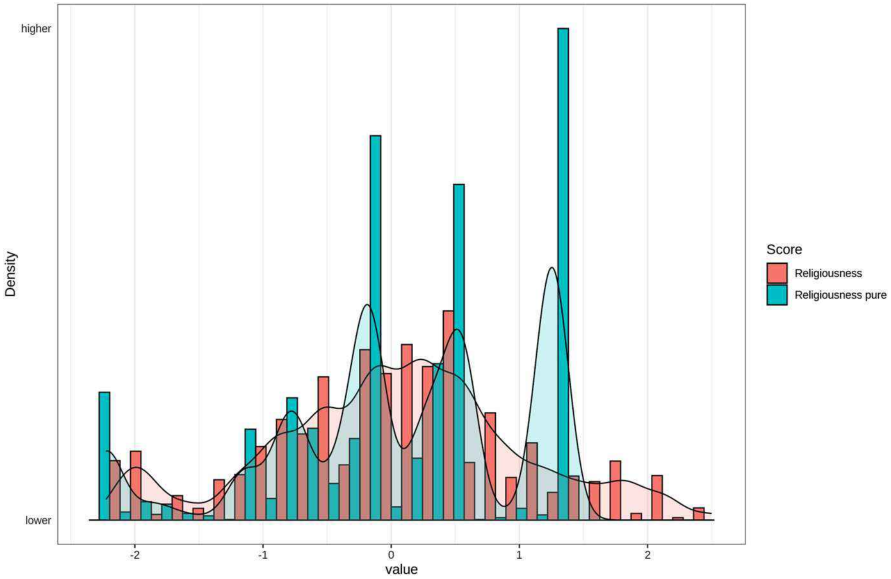

Fig. 3 Distribution of religiousness scores. Based on 12 and 6 items, respectively

there are also main effects of race, such that blacks are somewhat more religious than whites at the same level of intelligence (0.29 and 0.20 in the full/pure scale model, both p's < 0.001). This result is in line with many previous studies on this issue (e.g. Fitchett et al., 2007).

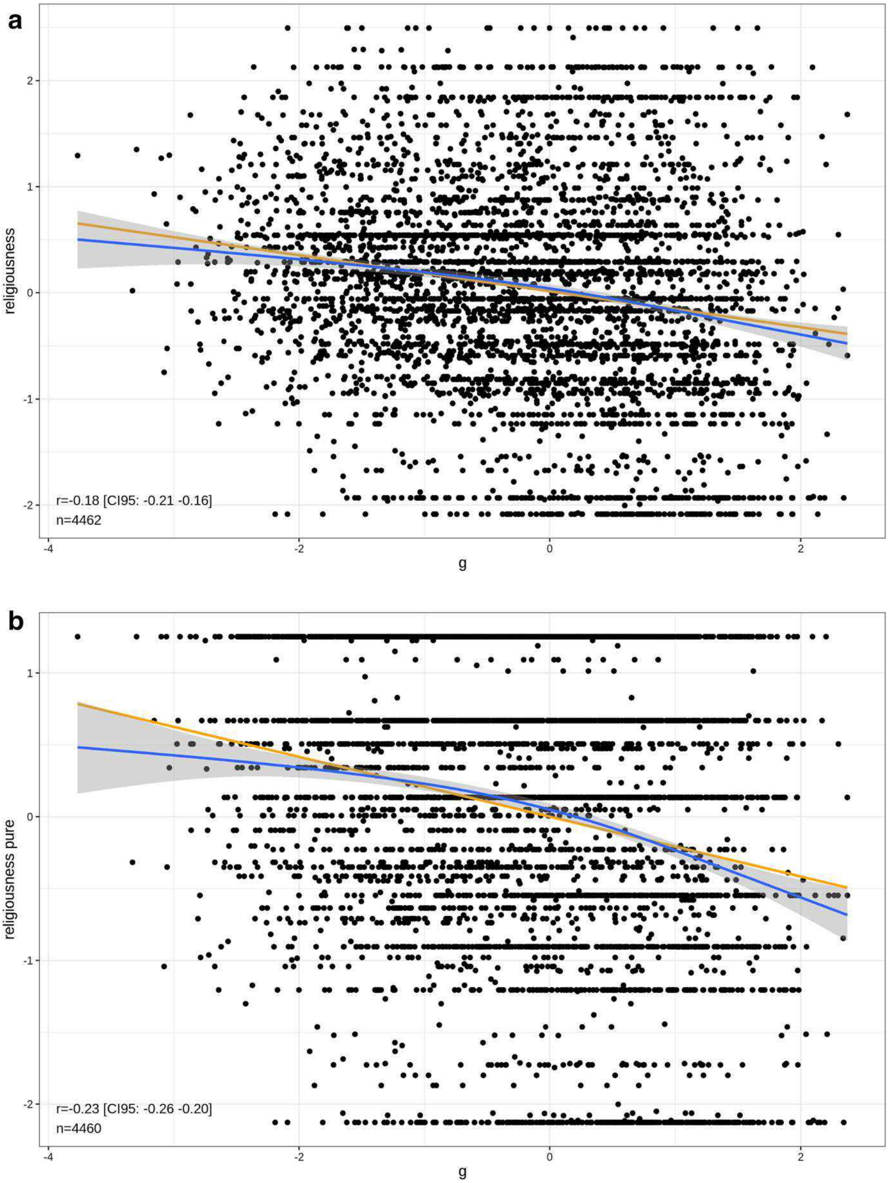

Fig. 4 **a**–**b** Scatterplot of intelligence (*g*) and religiousness as measured by 12 or 6 items. The orange line is the linear fit, and the blue line is the LOESS fit

We go further and analyze the item data as well. Item data were available for four of the cognitive tests (23 items from WAIS-INF, 6 from WAIS-BD, 112 from CVLT, and 52 from CFD; we excluded items with pass rates below 0.05 or above 0.95). We fit a single factor model using **mirt** for the items and saved the g-loadings.

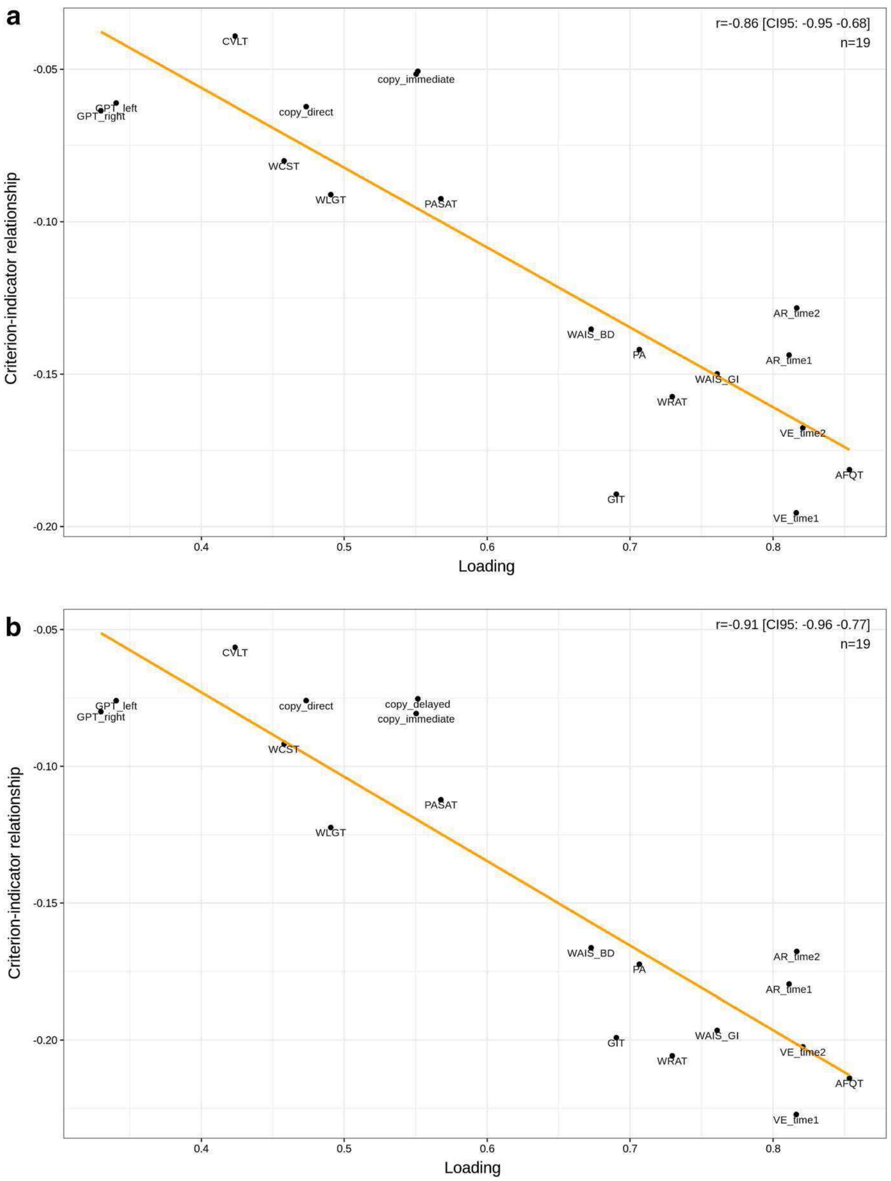

Fig. 5 **a**–**b** Jensen's method of correlated vectors used on the 19 cognitive tests for the total score (12 items, top) and pure scale (6 items, bottom)

Figure 7a, b shows the relationship between item g-loadings and religiousness (latent correlation, i.e. polychoric in this case; Uebersax, 2015).

The scatterplots revealed a strong outlying item from the WAIS Information scale. Religious people perform much better on this item than expected by their level of intelligence, suggesting that this item has some religious content. In fact, the item

Table 3 Regression model results for the full religiousness score (12 items)

Predictor/model	Basic	Demographic controls	Add interaction	Add SES	Basic whites	Full whites
Intercept	0.01 (0.015, 0.348)	− 0.38 (0.221, 0.089)	− 0.39 (0.221, 0.079)	− 0.51 (0.228, 0.027)	0.00 (0.016, 1)	− 0.64 (0.259, 0.013)
G	− 0.17 (0.014, < 0.001***)	− 0.15 (0.015, < 0.001***)	− 0.16 (0.016, < 0.001***)	− 0.14 (0.019, < 0.001***)	− 0.16 (0.016, < 0.001***)	− 0.15 (0.021, < 0.001***)
Age		0.01 (0.006, 0.088)	0.01 (0.006, 0.078)	0.01 (0.006, 0.026)		0.02 (0.007, 0.013)
Race = Black		0.15 (0.049, 0.002**)	0.29 (0.077, < 0.001***)	0.14 (0.050, 0.005**)		
Race = Hispanic		0.02 (0.071, 0.736)	0.07 (0.094, 0.44)	0.02 (0.072, 0.81)		
Race = Asian		− 0.23 (0.167, 0.165)	− 0.25 (0.169, 0.136)	− 0.22 (0.166, 0.186)		
Race = Native		0.00 (0.139, 0.973)	− 0.09 (0.150, 0.553)	− 0.02 (0.140, 0.896)		
*g** race = Black			0.12 (0.052, 0.019)			
*g** race = Hispanic			0.07 (0.078, 0.375)			
g * race = Asian			− 0.10 (0.146, 0.503)			
*g** race = Native			− 0.21 (0.131, 0.115)			
Education				0.01 (0.019, 0.696)		0.00 (0.021, 0.928)
Income				− 0.05 (0.016, 0.004**)		− 0.04 (0.018, 0.039)
R^2 adj	0.034	0.036	0.037	0.038	0.026	0.027
N	4462	4462	4462	4376	3654	3580

White race is the reference group

SES socioeconomic status

*$p < .01$; **$p < .005$; ***$p < .001$

Table 4 Regression model results for the pure religiousness scale (6 items)

Predictor/model	Basic	Demographic controls	Add interaction	Add SES	Basic whites	Full whites
Intercept	0.00 (0.015, 0.999)	−0.22 (0.219, 0.315)	−0.24 (0.219, 0.267)	−0.36 (0.225, 0.114)	0.00 (0.016, 0.999)	−0.45 (0.257, 0.081)
G	−0.21 (0.013, < 0.001***)	−0.20 (0.015, < 0.001***)	−0.22 (0.016, < 0.001***)	−0.18 (0.018, < 0.001***)	−0.22 (0.016, < 0.001***)	−0.19 (0.021, < 0.001***)
Age		0.01 (0.006, 0.314)	0.01 (0.006, 0.265)	0.01 (0.006, 0.114)		0.01 (0.007, 0.081)
Race = Black		0.05 (0.048, 0.35)	0.20 (0.076, 0.009*)	0.07 (0.050, 0.176)		
Race = Hispanic		−0.03 (0.070, 0.694)	0.07 (0.093, 0.454)	−0.01 (0.071, 0.868)		
Race = Asian		−0.45 (0.164, 0.007*)	−0.43 (0.167, 0.01*)	−0.43 (0.165, 0.009*)		
race = Native		0.00 (0.137, 0.994)	−0.09 (0.148, 0.562)	−0.03 (0.139, 0.853)		
*g** race = Black			0.14 (0.051, 0.008*)			
g * race = Hispanic			0.13 (0.077, 0.081)			
*g** race = Asian			0.09 (0.144, 0.515)			
*g** race = Native			−0.18 (0.129, 0.163)			
Education				−0.04 (0.018, 0.047)		−0.04 (0.021, 0.051)
Income				−0.02 (0.016, 0.321)		−0.01 (0.018, 0.568)
R^2 adj	0.052	0.053	0.054	0.054	0.048	0.048
N	4460	4460	4460	4374	3652	3578

White race is the reference group

SES socioeconomic status

$*p < .01$; $**p < .005$; $***p < .001$

asks “What is the main theme of the book of Genesis”? (the first book of the Bible). Aside from the outlier, the plots revealed a medium-sized negative correlation, in line with results from the test-level analysis in Figs. 5a, b. The relative weakness of the item-level results compared to the test-level results is perhaps best interpreted as being due to the increased sampling error in the estimates of the item statistics.

We then replicated our Jensen’s method results using standard approaches of examining for test bias. Specifically, for the test-level data, we employed local structural equation modelling (LSEM) to examine for measurement invariance for a continuous variable (moderator) (Hildebrandt et al., 2016). This is the continuous analogue of the more common multi-group confirmatory factor analysis approach (MGCFA) (Frisby & Beaujean, 2015; Lasker et al., 2019). We developed a model for the 19 tests in the battery. Since no prior theory existed on the topic, we used an iterative approach using the modification indexes and our professional judgement (Beaujean, 2014). The model was fit with the **lavaan** package (Rosseel et al., 2020), and we used the **sirt** package for LSEM (Robitzsch, 2020). The model was complex. We opted for a bi-factor approach with four group factors (verbal, memory, mathematics, and visual-spatial ability), two test occasion factors (time 1 and time 2), as well as 2 covariances. The appendix contains the details of this model. The model had excellent overall fit to the data RMSEA = 0.040, CFI = 0.985, TLI = 0.978, GFI = 0.977, SRM r = 0.025. We fit LSEM to the data in the region of − 1.5 to 1.5 with a bandwidth factor (h) of 5. Results revealed only minor differences in fit measures between high and low levels of religiousness, with slightly higher values for the religious. Figure 8 shows an example plot from this, and the full results can be found in the technical output. The modelled RMSEA was about 0.044 for persons with − 1.5 religiousness and 0.038 for those with 1.5. The other fit measures showed similar results.

For the items, we used differential item functioning (DIF) testing, as implemented in the **mirt** package. Specifically, we tested each cognitive item for differential functioning between each of the 12 religiousness items. We tested for both intercept and slope (discrimination) differences. We estimated the effect size of the bias using the approach advocated by (Meade, 2010) and implemented in the *empirical_ES()* function in **mirt**. Results are shown in Table 5.

All the DIF analyses found some items with differential functioning, but the directions were mixed, so the test-level effects were all near 0 (all were below 0.1 d). As a method check, we examined which item was most frequently detected as bias and found that this was WAIS Information item 18, the same item that was an outlier in the Jensen’s method analyses in Fig. 7a, b, again showing the congruence of results across methods.

We examined the different indicators of religiousness for differential relationships to intelligence. Specifically, we used Jensen’s method with the factor analysis results from the full scale. Results are shown in Fig. 9.

There was no detectable association between the religiousness factor loadings and the item’s relationship to intelligence (r = − 0.07). In other words, other factors than merely the association with overall religiousness were responsible for the association with intelligence at the item level. Examining the results, one can see that the items related to pure belief matters have stronger (negative) associations, while

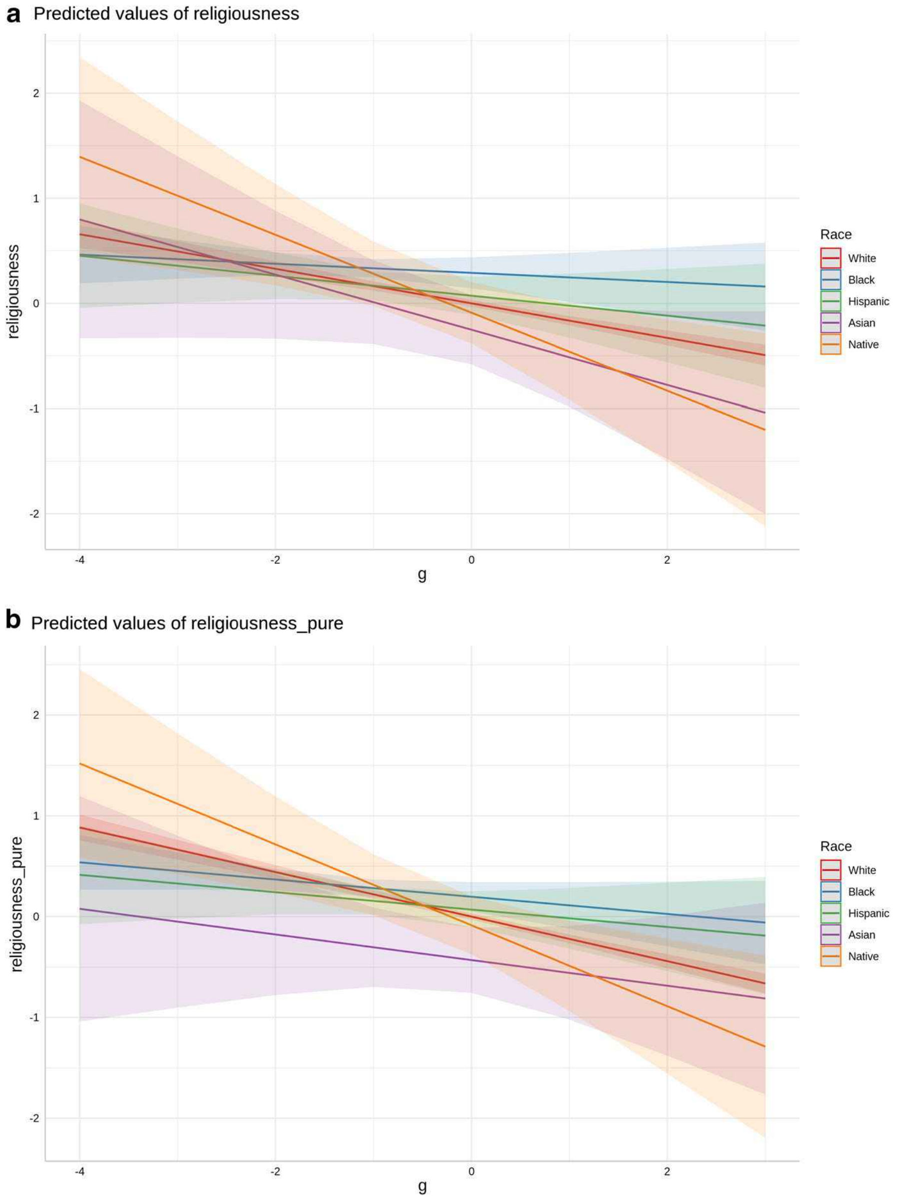

Fig. 6 **a**–**b** Regression slopes for intelligence on religiousness full scale (12 items, top), and pure scale (6 items, bottom), by race. Adjusted for age

those that involve other factors (impure) show near-zero associations, or even positive (church going). This diversity in associations was not due to demographic confounders that we previously examined (Tables 3, 4). When we fit regression models

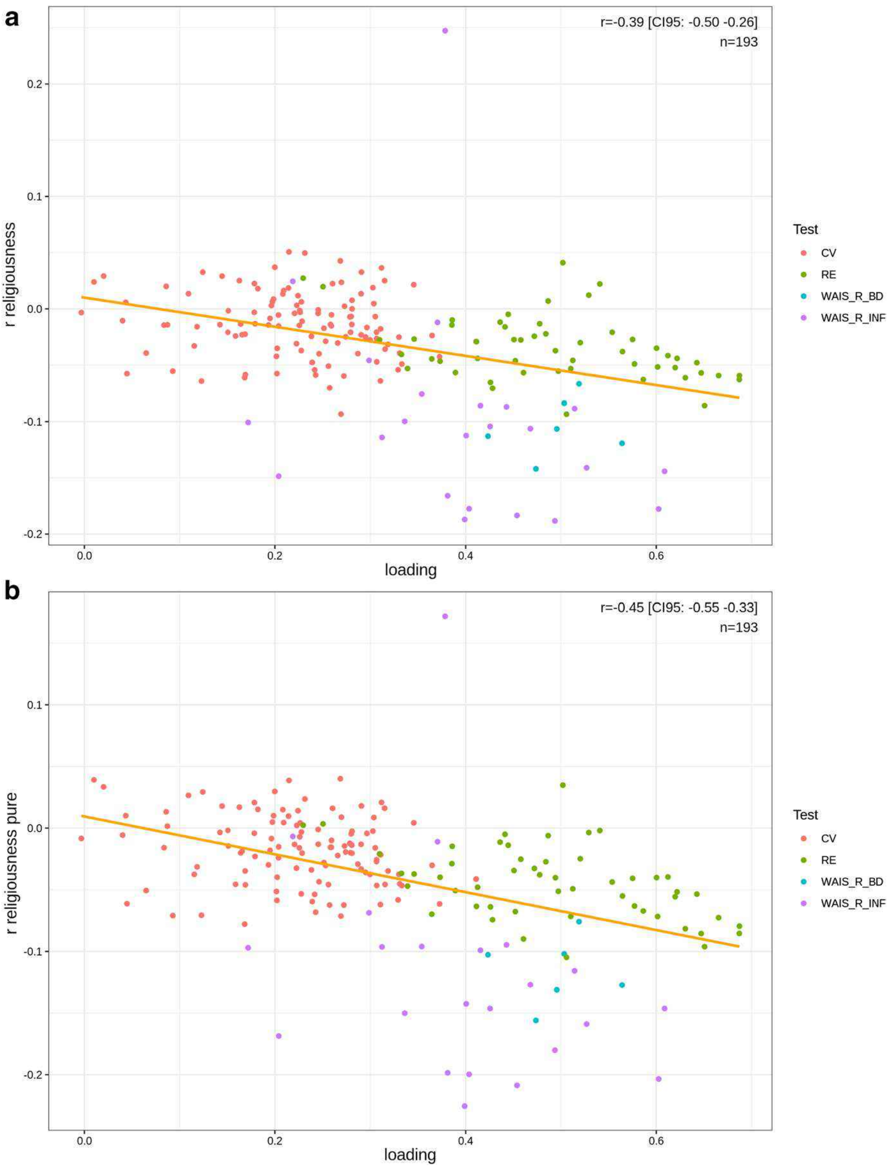

Fig. 7 **a**–**b** Scatterplot of item *g*-loadings and item correlations to religiousness total scale (12 items, top) and pure scale (6 items, bottom)

for each of the 12 religiousness items, we find the same patterns as shown in Fig. 8, results shown in Fig. 10.

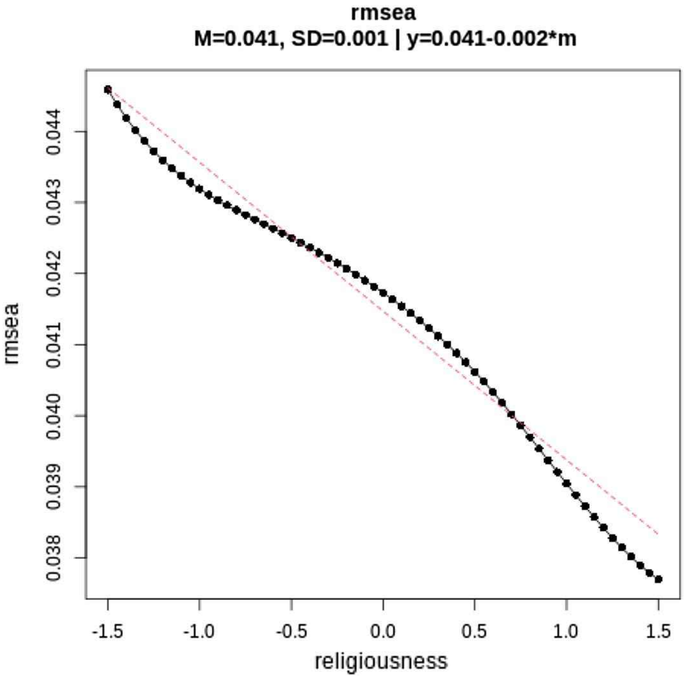

Fig. 8 LSEM results for religiousness (total scale, 12 items) for RMSEA measure of model fit

Table 5 Results of DIF analysis. Liberal refers to considering items as biased when $p < .05$, and conservative when this survived the Bonferroni correction

Test-level effect size, liberal	Test-level effect size, conservative	Bad items liberal	Bad items conservative
−0.02	0.01	20	1
−0.07	−0.05	56	17
0.01	0.01	27	3
0.00	0.01	23	1
0.00	0.01	10	2
−0.05	−0.05	49	13
−0.08	−0.06	43	9
−0.05	−0.06	68	17
−0.01	0.00	29	3
−0.02	−0.01	29	5
0.01	0.01	23	2
0.03	0.01	28	2

Effect size refers to the standardized (Cohen *d*) effect size at the test level for the partial fits with the offending items

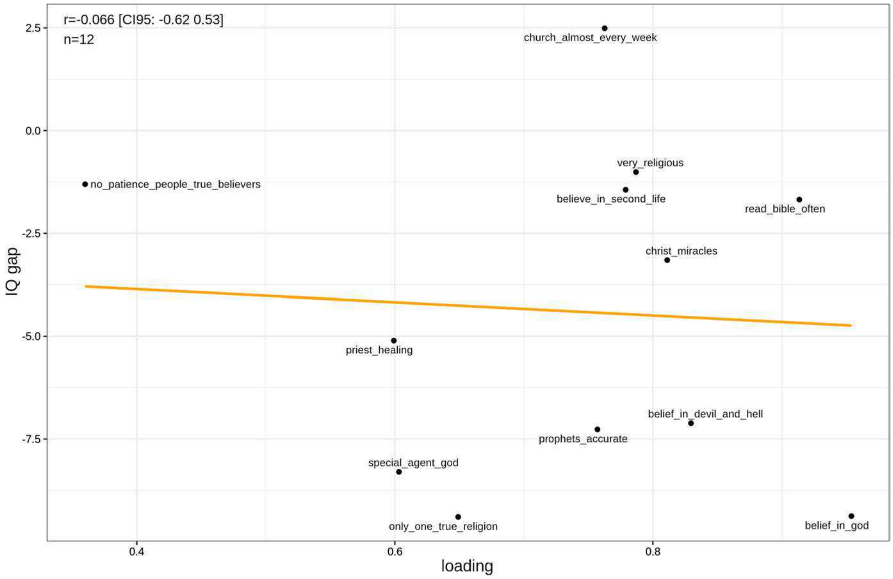

Fig. 9 Jensen's method applied to the 12 items from the total religiousness scale. The X axis is the item's factor loading on the religiousness factor. The Y axis is the IQ gap between persons who affirm and deny a given statement. One item was reversed due to negative factor loading (no patience for true believers).

Discussion

It appears, therefore, that this study refutes the findings presented in Dutton et al. (2020). It reaches very different results due to a different, and superior, method whereby (1) we have a large and representative individual-level data set (2) we are not dealing simply with church membership, which is an impure measure of religiosity, but rather with multiple measures of religiosity which we have also been able to factor analyse (3) we have a much larger number of intelligence tests and items allowing us to use Jensen's method, differential item functioning tests, and local structural equation modelling with a greater degree accuracy. All the methods agree on the finding that the religion-IQ nexus is principally concerned with *g* and thus one or more of theories presented above for this relationship may well explain it. It is notable here that Jensen's method (method of correlated vectors), which has been criticized for nonsensical results (Wicherts, 2017; Wicherts & Johnson, 2009), was actually congruent with the item response theory-based results, even detecting the same highly biased knowledge item. Thus, our study indirectly shows that this method can produce sensible results with item data, as long as the analysis is done correctly using item response theory-based metrics (for another example, see Al-Bursan et al., 2018).[1]

[1] This approach was first tested by Kirkegaard (2016) in an unpublished note. However, the problems with the classical test theory approach and corrections were in fact outlined by Jensen himself (Jensen 1980, p.437 and p.445; Jensen & McGurk 1987).

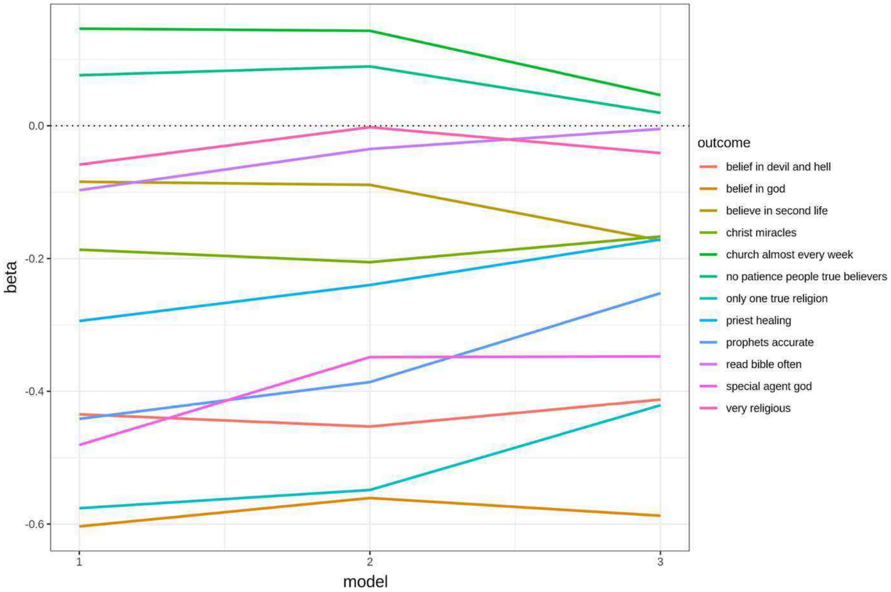

Fig. 10 Logistic regression results for 12 indicators of religiousness. Model 1 controlled for nothing (baseline), model 2 controls for age and race, and model 3 controls for age, race, income, and education

Many studies on this nexus have looked at a small number of measures of religiosity. The variety of religiosity measures we have examined here present us with a clear point of interest. Church going, in our dataset, is slightly positively associated with IQ, though other studies have found a weak negative association. One possible reason for this is that as religious belief and practice declines in Western countries, including America as has been widely documented (see Bruce, 2002), we would expect the relationship between indicators of not being religious—such as not believing in God or not going to church—to become more weakly associated with intelligence. The fact that some dimensions of church going—such as its civic and pro-social nature (see Jensen, 1998)—are positively correlated with intelligence would mean that at some point, this measure would become non-associated and eventually positively associated with intelligence. This would be strengthened by extreme liberal Christians who might not really literally believe in God but might attend church for assorted psychological reasons. In the Church of England, for example, there are certainly worshippers, and even priests, who do not seem to believe in God (see Freeman, 1993). On the other hands, if people live in communities in which the local church is highly influential then they may attend church despite not believing in God in order to avoid ostracism or to ensure social approval, something known as extrinsic religiousness (Hills et al., 2004). The other measures of religiosity which were only very weakly negatively correlated with intelligence, however, were all issues which would be consistent with being a liberal Christian. A liberal Christian may well read the Bible frequently. Indeed, reading frequently at all could be a reflection of intelligence, as smarter people read more in general (Ritchie et al., 2015). A liberal Christian may well have vague belief in the afterlife and they may even regard themselves as 'very religious', no matter what other people might think about them, if being 'religious' were a significant component of their identity in a relatively secular area. By contrast, rural fundamentalist Americans might not regard themselves as 'very religious' because everybody they know is, by ordinary standards, 'very religious'. Indeed, to complicate matters further, many fundamentalist Protestants insist that they are not 'religious' at all. Roman Catholics are 'religious', due to their perceived focus on ritual, but fundamentalist Protestants like themselves, by contrast, are 'Christian' (see Dutton, 2008). Thus, ironically, self-describing as 'very religious' in a US dataset may actually indicate that you are moderately religious.

The measures that were most strongly negatively associated with IQ were all markers of fundamentalism, such as belief in the Devil or in there only being one true religion. Intelligence is negatively associated with many of the markers of fundamentalism beyond mere religiosity including dogmatism, conservatism, and authoritarianism (Onraet et al., 2015), in that such churches promote strict obedience to authority (see Barr, 1977). In addition, belief in God among such people will be absolute. So these relationships make sense in terms of Nyborg's (2009) finding that in America, extreme liberal Christian churches have the highest average IQ—even higher than atheists, as these will include some political ideologues and extremists with these traits being negatively associated with IQ—and the most fundamentalist churches have the lowest average IQs. So, this is congruous with Nyborg's model whereby, overall, we all need a way to make sense of our world and those who lack the intelligence to be able to do so using science will turn towards religion. Kirkegaard and Lasker (2020) replicated Nyborg's thesis using a large sample of dating users, finding that within every religious group with sufficient sample size, the least religious were the highest in intelligence. Consistent with Dutton and Van der Linden, they may also be less instinctive, due to their intelligence, meaning that their cognitive bias towards religiosity is lower. More research into the causes of this relationship would be fruitful, but we hope we have contributed here by comprehensively demonstrating that is really a matter of general intelligence.

Study Limitations

The study had a variety of limitations. First, perhaps most notably, all subjects were male. It is possible that intelligence may relate differently to religiousness in women. In the meta-analysis by Zuckerman et al. (2020), this was tested and found not to be the case in several large datasets. As such, this limitation is not a significant concern.

Second, the measures of religiousness were self-reported in a psychiatric questionnaire. It is conceivable that this might distort the results, depending on any self-report biases involved in filling this out. We are not aware of this evidence this may be the case, however. We are also not aware of any study that used other-reported religiousness and intelligence, to see if the mode of reporting may affect the patterns.

Third, the study was carried out many years ago, with the second wave being collected in 1985–1986. The USA has then seen a large increase in atheism rates, whereas it has historically ‘halted behind’ other European-descended countries (Pew Research Center, 2019). Possibly this change in the distribution of religiousness has affected the associations with intelligence.

Funding No funding was supplied for this research.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Al-Bursan, I. S., Kirkegaard, E. O., Fuerst, J., Bakhiet, S. F. A., Al Qudah, M. F., Hassan, E. M. A. H., & Abduljabbar, A. S. (2018). Sex differences in 32,347 Jordanian 4th graders on the National Exam of Mathematics. *Journal of Individual Differences, 40*(2), 71–81. https://doi.org/10.1027/1614-0001/a000278

Barr, J. (1977). *Fundamentalism*. SCM.

Bayroff, A.G., & Fuchs, E.F. (1970). *The armed services vocational aptitude battery*. U. S. Army Behavior and Systems Research Laboratory. https://apps.dtic.mil/docs/citations/AD0706832.

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide* (1st ed.). Routledge.

Bruce, S. (2002). *God is dead: Secularisation in the West*. Blackwell.

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2020). *mirt: Multidimensional Item Response Theory* (1.32.1) [Computer software]. https://CRAN.R-project.org/package=mirt.

Dahlstrom, W., Welsh, G., & Dahlstrom, L. (1975). *An MMPI handbook volume I clinical interpretation a* (revised). Minnesota: University of Minnesota Press.

Dutton, E. (2014). *Religion and intelligence: An evolutionary analysis*. Ulster Institute for Social Research.

Dutton, E. (2008). *Meeting Jesus at University: Rites of passage and student evangelicals*. Ashgate.

Dutton, E., & Van der Linden, D. (2017). Why is intelligence negatively associated with religiousness? *Evolutionary Psychological Science, 3*, 392–403. https://doi.org/10.1007/s40806-017-0101-0

Dutton, E., te Nijenhuis, J., Madison, G., Van der Linden, D., & Metzen, D. (2020). The myth of the stupid believer: The negative religiousness-IQ Nexus is not on general intelligence (*g*). *Journal of Religion and Health, 59*, 1567–1579. https://doi.org/10.1007/s10943-019-00926-3

Elwood, R. W. (1995). The California verbal learning test: Psychometric characteristics and clinical application. *Neuropsychology Review, 5*(3), 173–201. https://doi.org/10.1007/BF02214761

Freeman, A. (1993). *God in us: A case for Christian humanism*. SPCK.

Frisby, C. L., & Beaujean, A. A. (2015). Testing Spearman’s hypotheses using a bi-factor model with WAIS-IV/WMS-IV standardization data. *Intelligence, 51*, 79–97. https://doi.org/10.1016/j.intell.2015.04.007

Gilkey, C. (1924). Religion among American students. *Journal of Religion, 1*, 4.

Greve, K. W., Stickle, T. R., Love, J. M., Bianchini, K. J., & Stanford, M. S. (2005). Latent structure of the Wisconsin Card Sorting Test: A confirmatory factor analytic study. *Archives of Clinical Neuropsychology, 20*(3), 355–364. https://doi.org/10.1016/j.acn.2004.09.004

Fitchett, G., Murphy, P., Kravitz, H., Everson-Rose, S., Krause, N., & Powell, L. (2007). Racial/ethnic differences in religious involvement in a multi-ethnic cohort of midlife women. *Journal for the Scientific Study of Religion, 46*, 119–132. https://doi.org/10.1111/j.1468-5906.2007.00344.x

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research, 51*(2–3), 257–258. https://doi.org/10.1080/00273171.2016.1142856

Hills, P., Francis, L., Argyle, M., & Jackson, C. (2004). Primary personality trait correlates of religious practice and orientation. *Personality and Individual Differences, 36*, 61–73. https://doi.org/10.1016/S0191-8869(03)00051-5

Howells, T. (1928). A comparative study of those who accept as against those who reject religious authority. *University of Iowa Studies in Character, 2*, 3.

Jensen, A. R. (1980). *Bias in mental testing*. Free Press.

Jensen, A. R., & McGurk, F. C. J. (1987). Black-white bias in 'cultural' and 'noncultural' test items. *Personality and Individual Differences, 8*, 295–301. https://doi.org/10.1016/0191-8869(87)90029-8

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Kanazawa, S. (2012). *The intelligence paradox: Why the intelligent choice isn't always the smart one*. Wiley.

Kirkegaard, E. O. W. (2016). *Using classical test theory statistics with Jensen's method*. R Notebook https://rpubs.com/EmilOWK/230077.

Kirkegaard, E. O. W., & Lasker, J. (2020). Intelligence and Religiosity among Dating Site Users. *Psych, 2*, 25–33. https://doi.org/10.3390/psych2010003

Kirkegaard, E. O. W., & Nyborg, H. (2020). Pupil size and intelligence: A large-scale replication study. *Mankind Quarterly, 60*(4), 525–538.

Lasker, J., Pesta, B. J., Fuerst, J. G. R., & Kirkegaard, E. O. W. (2019). Global ancestry and cognitive ability. *Psych, 1*(1), 431–459. https://doi.org/10.3390/psych1010034

Leckliter, I. N., Matarazzo, J. D., & Silverstein, A. B. (1986). A literature review of factor analytic studies of the WAIS-R. *Journal of Clinical Psychology, 42*(2), 332–342. https://doi.org/10.1002/1097-4679(198603)42:2%3c332::AID-JCLP2270420220%3e3.0.CO;2-2

Lynn, R., Harvey, J., & Nyborg, H. (2009). Average intelligence predicts atheism rates across 137 nations. *Intelligence, 37*(1), 11–15. https://doi.org/10.1016/j.intell.2008.03.004

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95*(4), 728–743. https://doi.org/10.1037/a0018966

Meisenberg, G., Rindermann, H., Patel, H., & Woodley, M. A. (2012). Is it smart to believe in God? The relationship of religiosity with education and intelligence. *Temas Em Psicologia, 20*, 101–120.

Nyborg, H. (2009). The intelligence-religiosity nexus: A representative study of white adolescent Americans. *Intelligence, 37*, 81–93. https://doi.org/10.1016/j.intell.2008.08.003

Nyborg, H., & Jensen, A. R. (2000). Black–white differences on various psychometric tests: Spearman's hypothesis tested on American armed services veterans. *Personality and Individual Differences, 28*(3), 593–599. https://doi.org/10.1016/S0191-8869(99)00122-1

Nyborg, H., & Jensen, A. R. (2001). Occupation and income related to psychometric *g*. *Intelligence, 29*(1), 45–55. https://doi.org/10.1016/S0160-2896(00)00042-8

Onraet, E., Van Hiel, A., Dhont, K., Hodson, G., Schittekatte, M., & De Pauw, S. (2015). The association of cognitive ability with right–wing ideological attitudes and prejudice: A meta–analytic review. *European Journal of Personality, 29*(6), 599–621. https://doi.org/10.1002/per.2027

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2016). Atheists and agnostics are more reflective than religious believers: Four empirical studies and a meta-analysis. *PLoS ONE, 11*(4), e0153039. https://doi.org/10.1371/journal.pone.0153039

Pew Research Center. (2019). In U.S., decline of Christianity continues at rapid pace. *Pew Research Center's Religion and Public Life Project*. https://www.pewforum.org/2019/10/17/in-u-s-decline-of-christianity-continues-at-rapid-pace/.

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research* (1.9.12.31) [Computer software]. https://CRAN.R-project.org/package=psych.

Ritchie, S. J., Bates, T. C., & Plomin, R. (2015). Does learning to read improve intelligence? A longitudinal multivariate analysis in identical twins from age 7 to 16. *Child Development, 86*(1), 23–36. https://doi.org/10.1111/cdev.12272

Rindermann, H., Flores-Mendoza, C., & Woodley, M. A. (2012). Political orientations, intelligence and education. *Intelligence, 40*, 217–225. https://doi.org/10.1016/j.intell.2011.11.005

Robitzsch, A. (2020). *sirt: Supplementary item response theory models* (3.9-4) [Computer software]. https://CRAN.R-project.org/package=sirt.

Rosseel, Y., Jorgensen, T. D., Rockwood, N., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., Hallquist, M., Rhemtulla, M., Katsikatsou, M., Barendse, M., & Scharf, F. (2020). *lavaan: Latent variable analysis* (0.6-7) [Computer software]. https://CRAN.R-project.org/package=lavaan.

Ruff, R. M., & Parker, S. B. (1993). Gender- and age-specific changes in motor speed and eye-hand coordination in adults: Normative values for the finger tapping and grooved pegboard tests.

Perceptual and Motor Skills, 76(3_suppl), 1219–1230. https://doi.org/10.2466/pms.1993.76.3c.1219

Sherkat, D. (2002). Sexuality and religious commitment in the United States: An empirical examination. *Journal for the Scientific Study of Religion, 41*, 313–323. https://doi.org/10.1111/1468-5906.00119

Shin, M.-S., Park, S.-Y., Park, S.-R., Seol, S.-H., & Kwon, J. S. (2006). Clinical and empirical applications of the Rey–Osterrieth Complex Figure Test. *Nature Protocols, 1*(2), 892–899. https://doi.org/10.1038/nprot.2006.115

Sorjonen, K., Aurell, J., & Melin, B. (2017). Predicting group differences from the correlation of vectors. *Intelligence, 64*, 67–70. https://doi.org/10.1016/j.intell.2017.07.008

Templ, M., Alfons, A., Kowarik, A., & Prantner, B. (2015). *VIM: Visualization and imputation of missing values*. CRAN. http://cran.r-project.org/web/packages/VIM/index.html.

The Centers for Disease Control Vietnam Experience Study. (1988a). Health Status of Vietnam Veterans: I. Psychosocial Characteristics. *JAMA, 259*(18), 2701–2707. https://doi.org/10.1001/jama.1988.03720180027028

The Centers for Disease Control Vietnam Experience Study. (1988b). Health Status of Vietnam Veterans: II. Physical Health. *JAMA, 259*(18), 2708–2714. https://doi.org/10.1001/jama.1988.03720180034029

The Centers for Disease Control Vietnam Experience Study. (1988c). Health Status of Vietnam Veterans: III. Reproductive Outcomes and Child Health. *JAMA, 259*(18), 2715–2719. https://doi.org/10.1001/jama.1988.03720180041030

Tombaugh, T. N. (2006). A comprehensive review of the Paced Auditory Serial Addition Test (PASAT). *Archives of Clinical Neuropsychology, 21*(1), 53–76. https://doi.org/10.1016/j.acn.2005.07.006

Uebersax, J. S. (2015). *Introduction to the tetrachoric and polychoric correlation coefficients*. http://john-uebersax.com/stat/tetra.htm.

Wicherts, J. M., & Johnson, W. (2009). Group differences in the heritability of items and test scores. *Proceedings of the Royal Society B: Biological Sciences, 276*(1667), 2675–2683. https://doi.org/10.1098/rspb.2009.0238

Wicherts, J. M. (2017). Psychometric problems with the method of correlated vectors applied to item scores (including some nonsensical results). *Intelligence, 60*, 26–38. https://doi.org/10.1016/j.intell.2016.11.002

Witt, J. C. (1986). Review of the wide range achievement test-revised. *Journal of Psychoeducational Assessment, 4*(1), 87–90. https://doi.org/10.1177/073428298600400110

Zuckerman, M., Li, C., Lin, S., & Hall, J. (2020). The negative intelligence–religiosity relation: New and confirming evidence. *Personal and Social Psychology Bulletin, 46*, 856–868. https://doi.org/10.1177/0146167219879122

Zuckerman, M., Silverman, J., & Hall, J. (2013). The relation between intelligence and religiosity: A meta-analysis and some proposed explanations. *Personality and Social Psychology Review, 17*, 325–354. https://doi.org/10.1177/1088868313497266

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

View publication stats