Table of contents

Introduction

In this project, I am going to analyze the data obtained from the openpsychometrics.org website, where any users can fill out several personality questionnaires. I use data from the Humor Styles Questionnaire (Martin, 2003). This questionnaire was developed due to no other self-report measurements being available at the time that would quantify humor styles. The main purpose of the assessment was to distinguish between healthy and unhealthy humor styles – that is, which kind of humor contributes to individuals' well-being and which humor styles are malicious for the person, but also the surroundings. After examining literature on humor, the authors conclude humor is either used to enhance one self, or to enhance relationships with others (e.g. using humor as a mechanism to cope with everyday life or strengthening relationships with others through humor). With this framework, a 2x2 format was adapted. Based on this theoretical background, four humor styles were established – the Affiliative humor (AF) scale, Self-enhancing humor (SE) scale, Aggressive humor (AG) scale and Self-defeating humor (SD) scale. The AF style is a healthy way of making others laugh, telling jokes often and being a person perceived as having a sense of humor. On the other hand, the AG humor scale is often aimed towards others, putting them down, degrading them and often insulting them under the disguise of humor. The SE scale is a healthy way to enhance oneself – through a positive outlook on life and as a defense mechanism to help cope with negative feelings. Last scale, SD, is an unhealthy way of making others laugh and facilitating social relationships – it consists of putting yourself down, joking at your own expense over the line and seeking approval from others.

In the original paper, Martin et. al described the process behind creating the HSQ – several validation rounds with multiple samples starting with over 60 items decreased to overall 32 items (for full list of items, see Supplement materials), using both college student samples and a senior sample. The split into four factors was confirmed by a factor analysis fit by the authors with good fit (RMSEA = .048). The HSQ has good measures of reliability, internal consistency as well as validity and provides a tool to assess individual's dominant humor styles. In the initial study, Martin et. al used other measures, such as well-being assessment questionnaires or the NEO-PI-R personality assessment to confirm predictive validity of the measurement tool and its usefulness in practice. Since it's been published, the paper was cited over 500 times and the HSQ was translated into more than 30 languages and it has proven to have satisfying psychometric characteristics across language mutations, though with similar inconsistencies across studies (Heintz & Ruch, 2019).

My goal is to not only replicate some of the analyses in the original paper, but to provide a more detailed look into the questionnaire, define its weaknesses and recommend improvements to the item pool. As far as I know, there was no rigorous analysis of the items and studies usually focus on PCA, CFA, reliability measures and construct/predictive validity (Heintz & Ruch, 2016; Chen & Martin, 2007; Baughman et al., 2012). Thus, I want not only to validate the HSQ on yet another sample, but use more methods to assess the item pool and suggest improvements.

Methods

The dataset consists of 1071 responses to 32 statements. Each statement was rated on a 1 – 5 likert scale, based on the respondent's agreement with that statement (e.g.: "If I am feeling sad or upset, I usually lose my sense of humor."), where 1 = Never or very rarely true, 2 = Rarely true, 3 = Sometimes true, 4 = Often true, 5 = Very often or always true. If an item was left unanswered, it was

coded as "-1". Furthermore, each participant entered their gender (male, female or other) and age (any number). A total score was also calculated and while it is included in the raw dataset, it had to be removed for the data analysis, because an incorrect algorithm was used. Total scores were calculated simply as a sum of the answers to all questions on each scale (reverse coded items in mind).

Before actual analysis, the data had to be manipulated first. I marked all the missing responses as "NA" and since some participants left out more than one answer, I removed these participants completely to avoid issues with algorithms and possible biases (78 participants left out). Then, I had to re-code a few reverse scored items and make sure there was no more missing data or non-sense data (e.g. an answer "6" or "0"). I am left with a sample of 993 respondents which I am looking to analyze further. After I cleaned the dataset, I split the data into four sub-scales according to the theory – each sub-scale had 8 questions, which were non-randomly ranked as following: AF, SE, AG, SD, AF, SE, etc.

First, I am going to describe the sample via several descriptive measures (median, mean, SD, etc.) and look at the distributions of total scores for all scales. Then, I want to explore the frequency of answers on the likert scale (for the whole sample as well as individual scales). Afterwards, I want to assess validity via correlation matrixes between individual items for each scale, as well as correlations between scales. On top of that, I will be estimating reliability via several measures – two split-half methods and Cronbach's alpha coefficients. After that, I want to estimate item difficulties using the ordinal scale and discriminations using the Upper-Lower Index (ULI) to compare the first and third groups. I will also calculate the RIT and RIR discrimination coefficients and Cronbach's alpha drop upon item removal. Then, I will assess what polytomous and dichotomous (with binarized data) model to use and fit them accordingly. I will also plot a Wright map for selected scale and use IRT detection methods to identify differentially functioning items. Afterwards, I will fit a factor analysis, assess goodness of fit and look for any differences between groups (based on age and gender).

Results

Descriptive statistics

The data was collected via an online survey on openpsychometrics.org, therefore only for English speaking participants. In the sample were 537 male participants, 443 female participants, 8 "other gender" participants and 5 participants who did not specify gender. Age was also fulfilled, ranging from 14 to 70 (*Mean = 26, SD = 10.8, median = 22*). Four people submitted assumed false age (above 100), those were included in most analyses, but excluded in the analyses in which age was used as a variable.

First, I decided to look at the descriptive statistics of the sample. As mentioned in the paper by Martin, 2003, the AF scale had a larger total score mean – the reason is assumed to be the universality of the scale and its questions, e.g. "I laugh and joke a lot with my closest friends." – the question is; who doesn't laugh a lot with their close friends? This is confirmed in the presented sample as well, as the AF scale is different in terms of descriptive statistics of the total score (see Table 1). The SD is also slightly smaller, which may amplify the fact that most people score high on this scale. This is consistent with Martin et. al, though they obtained a larger mean (46.4) for the AF scale. AG/SD scales have the lowest mean total scores, corresponding with the original study. I also included a percentile estimate for each scale and participant based on the sample.

Table 1 – Descriptive statistics for all scales

	Minimum	Maximum	Mean	Median	SD
Affiliative humor	10	40	32.015	33	5.617
Self-Enhancing humor	8	40	26.987	27	6.022
Aggressive humor	8	40	23.002	23	6.235
Self-Defeating humor	8	40	21.755	21	6.312

It is also interesting to look at the total score distributions on each scale. I tested the normality of the distributions using the Shapiro-Wilk's test and for all the four scales, null-hypothesis was rejected and therefore, none of the distributions are normal on a .05 significance level. However, it is important to note that the sample size is quite large and even small differences may result in rejecting the null hypothesis of normality. As can be seen in Figure 1 though, the AF scale is very far from a normal distribution, the total score averages clearly lean towards higher numbers. When we also look at Figure 2, we can see that more than 40% of all answers on the AF scale were "5", or "Very often or always true". Compared to the distribution of answers for the whole test (Figure 3), it's obvious that the AF scale differs greatly in terms of participant responses. The rest may lean slightly to the lower or higher end, but more-or-less stays in the middle. In the R code (Supplement materials), histograms plotting answer frequency for each item alone can be found.

Figure 1 – Total score distributions on each scale

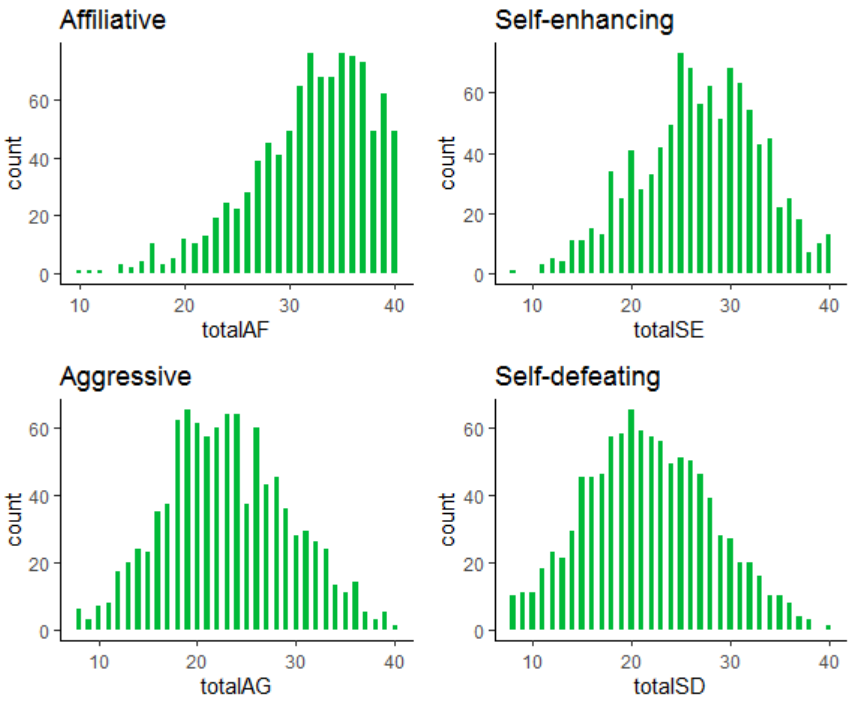


Figure 2 – Frequency of answers on each scale

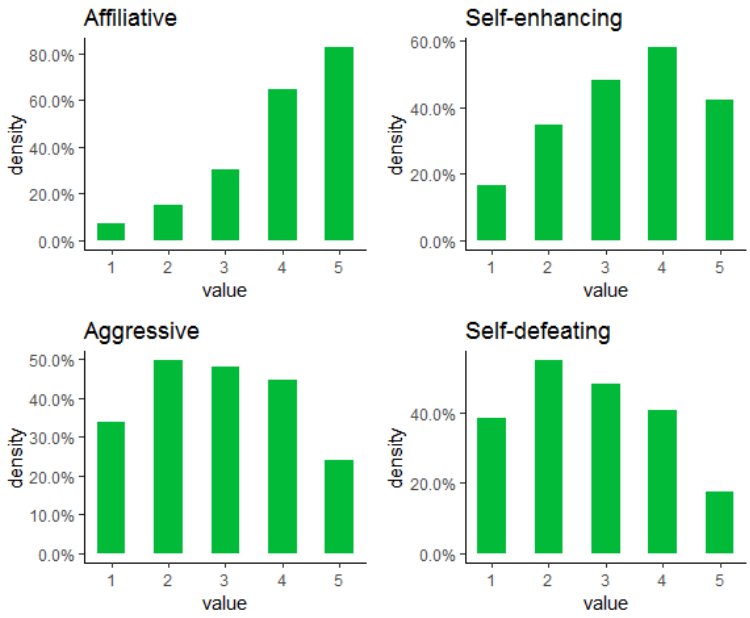
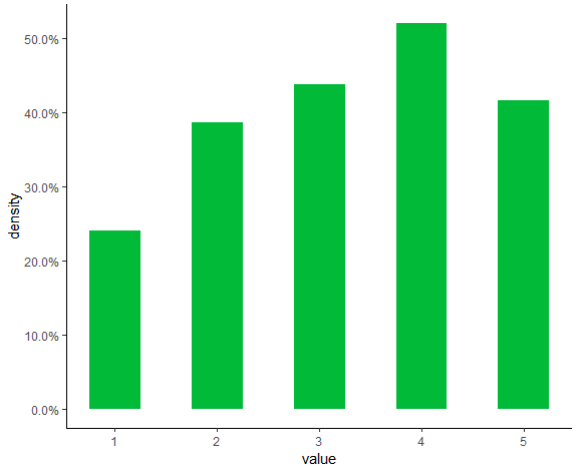


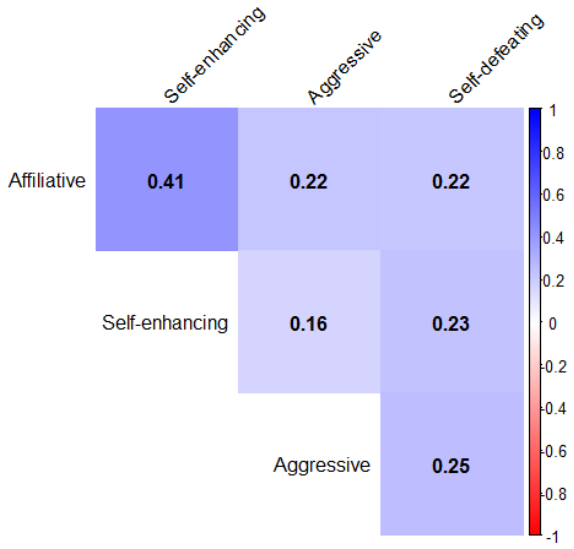
Figure 3 - Frequency of answers on the whole test



Internal consistency

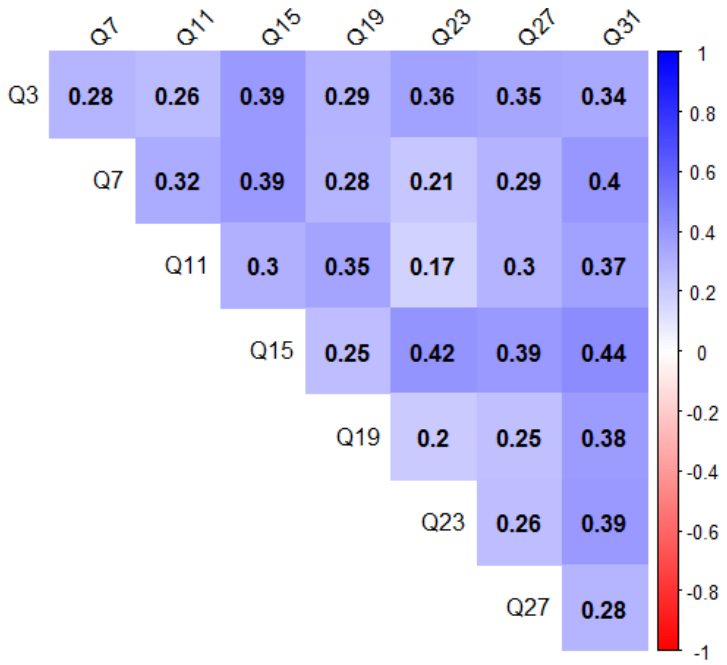
First, let me present a correlation matrix between the four scales (Figure 4). Significance is marked as colored – any insignificant correlations will be white squares (significance level = .01). As can be seen in Figure 4, both the healthy scales (AF and SE) are correlated ($r = .41$) and the second largest correlation is between both unhealthy scales (SD and AG) ($r = .25$) – this is consistent with Martin et. al findings, though they obtained smaller, insignificant correlations between scales that theoretically should not correlate – e.g. SE and AG. That said, the authors originally differentiated between males and females and it is entirely possible this difference is due to this fact.

Figure 4 - Correlations between scales



Considering that even according to further analyses, the AF scale is vastly different from the rest, I will mostly comment on the Aggressive (AG) scale in this section – correlation matrixes for other scales are available in the R code (Supplement materials). I chose the aggressive scale since the whole purpose of the questionnaire is to assess unhealthy styles of humor. For that very reason, I think it's best to have this scale as refined as possible. As for the internal consistency, below can be found the correlation matrix for items for the AG scale (Figure 5). Again, all correlations are significant ($p < .01$) – possibly caused by the size of the sample.

Figure 5 - Correlation matrix for the AG scale items



The largest correlation between Q31 and Q15 (.44) correspond with the wording of these items – "*Even if something is really funny to me, I will not laugh or joke about it if someone will be offended*", "*I do not like it when people use humor as a way of criticizing or putting someone down*".

Someone with a high preference of AG humor will most likely respond similarly to these questions, as they both assess the intention of the individual to offending others or putting them down with their humor.

The smallest correlation (.17) is between items Q11 and Q23 (respectively): *"I never participate in laughing at others even if all my friends are doing it"* and *"When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it."* It honestly does not make much sense to me as to why these two items are correlated so low, it is probably due to a different reason than the wording (item Q11 was re-coded correctly as reverse scored).

Reliability

To assess reliability of the HSQ, I used split-half methods and Cronbach's alpha. Because of the inability to test-retest participants, this was the best suitable method to estimate how reliable the questionnaire is. First, I used a random split-half method to assess each scale' reliability – except the AG (*r = .62*), other scales were correlated r > .71. However, after using the Spearman-Brown formula to adjust for test length, all three scales were correlated > .83 (except for AG, *r = .77*). I decided to then use the test-retest method that calculates all possible split halves and then provides a mean reliability coefficient. With this method, all reliability coefficients scored above the satisfactory mark of reliability and thus proved all scales to be reliable (see Table 2). If we look at internal consistency measured by Cronbach's alpha, we see that all four scales – though the AG scale is again slightly less consistent – show satisfactory levels of the alpha coefficient. This finding is consistent with other studies in which the AG scale had significantly lower reliability and internal consistency than other scales (Baughman et al., 2012; Chen & Martin, 2007).

Table 2 - Reliability estimates for all scales

	Random split-half	All split-halves	Cronbach's alpha
Affiliative	0.8418495	0.8474419	0.8394266
Self-enhancing	0.8332573	0.8206417	0.8217907
Aggressive	0.7672811	0.7895528	0.7899427
Self-defeating	0.8434178	0.8219492	0.8193314

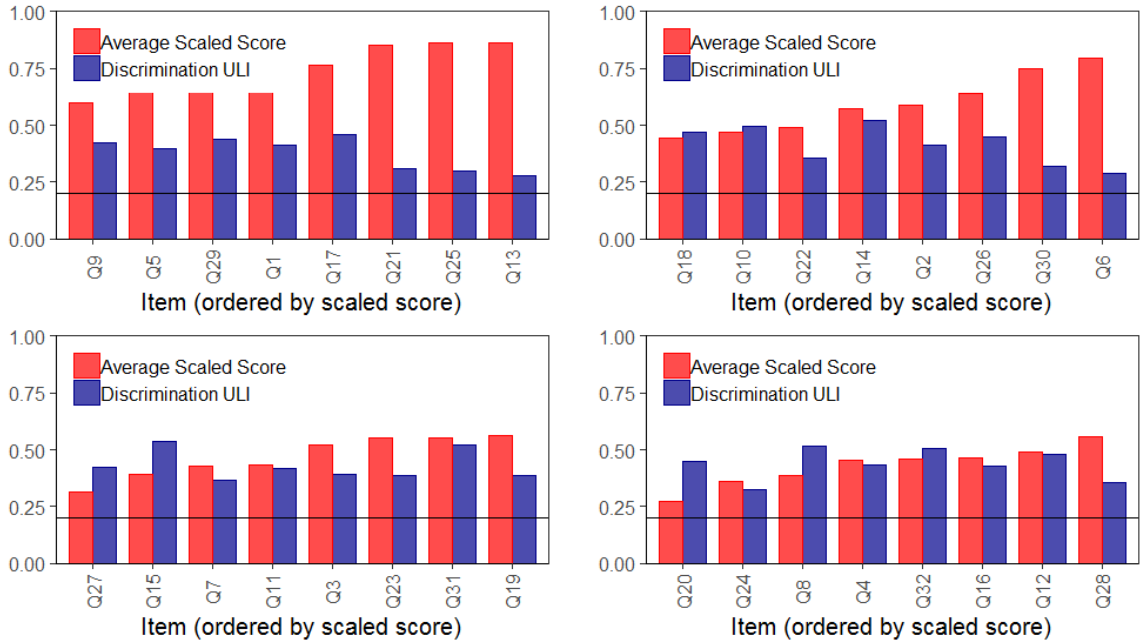
Traditional item analysis

For estimating item difficulties and discriminations, I decided not to binarize the data, because it is not clear whether to assign the answer 3 – "sometimes true" as agreement or disagreement with the statement. Later with binarized IRT models, I decided to code this answer as "0", due to the fact we want to identify prevalent humor styles and "sometimes" is not quite assessing prevalence well. There is also little reason to lose information by binarizing the data. For discrimination, I used the Upper-Lower Index (ULI), comparing the first and third group.

As can be seen in Figure 6 below, the AF scale is again vastly different from the rest. Most statements are not difficult (which is not a good term for psychological tests – difficulty in this case means how many people identify with that statement to be true) and the three items with least

difficulty – Q21 ("*I enjoy making people laugh*"), Q13 ("*I laugh and joke a lot with my closest friends*") and Q25 ("*I don't often joke around with my friends*") – also have very little discrimination. Just by looking at the wording of these items, it seems like these statements would be true for a very large portion of people, which accounts for the very large difficulty and goes hand in hand with the low discrimination. If most people answer positively to this statement – even those who do not have high scores on the AF scale otherwise – how is the item supposed to discriminate well? It is perhaps rare that people do not actually enjoy making other people laugh. That said, we need to keep in mind the fact that it is entirely possible we have a sample which identifies with the AF scale well, or that overall most humans tend to identify with the AF scale. It is difficult to assess at this point whether the items are poorly worded, or whether we are dealing with a unique sample/overall human tendency.

Figure 6 - Discrimination & difficulty plost for all scales (top left to bottom right - AF, SE, AG, SD scales)



The hypothesis of general human tendency towards the preference "healthy" AF and SE styles is further strengthened by the fact that the AG and SD styles have lower difficulties for all items, implying that it is possible "unhealthy" humor styles are not that prevalent. They do, however, have seemingly satisfying discrimination, which means that they do discriminate well between respondents who do and do not prefer these styles of humor. Let's look more closely at the AG scale (Table 3). The item with least difficulty (Q27: *"If I don't like someone, I often use humor or teasing to put them down*") is possibly because this hostile behavior is very seldom in human interaction. Putting down other people is an undesirable type of behavior that most people do not approve of. On the other hand, the item with highest difficulty (Q19: *"Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation"*) may apply even to people who are generally not fond of the AG humor style with no intention to hurt other people or put them down (for a traditional item analysis table with items from all scales, please see Supplement materials). However, it seems like all items (if removed) would still decrease Cronbach's alpha of the overall test. The exception to this is the item Q19, which if removed would not decrease but neither increase this coefficient. If we look at the wording, it is most likely due to the fact this does not accurately reflect the AG humor style, but an overall tendency of people saying funny things

in inappropriate situations without the intention to hurt others. I suggest rewording this item – if the intention to hurt was worked into the statement, it could more accurately reflect this style of humor.

Table 3 - Traditional item analysis for the AG scale

	Scale	Difficulty	SD	Discrimination ULI	Discrimination RIT	Discrimination RIR	Alpha Drop	Customized Discrimination
Q3	AG	3.08	1.16	1.57	0.64	0.51	0.77	0.39
Q7	AG	2.71	1.08	1.47	0.61	0.49	0.77	0.37
Q11	AG	2.73	1.24	1.68	0.61	0.46	0.77	0.42
Q15	AG	2.57	1.34	2.14	0.72	0.58	0.75	0.53
Q19	AG	3.24	1.23	1.55	0.59	0.44	0.78	0.39
Q23	AG	3.20	1.18	1.54	0.59	0.44	0.77	0.39
Q27	AG	2.26	1.26	1.69	0.62	0.47	0.77	0.42
Q31	AG	3.21	1.28	2.09	0.71	0.59	0.75	0.52

Polytomous models

Since the data is ordinal with polytomous answers on a likert scale, I wanted to fit a polytomous IRT model first to try to keep the most information possible. In the next section, I also binarized the data and used a dichotomous model.

First, I decided to test what model would be best to use for the AG scale which I will focus on (R code for other scales can be found in the Supplement materials). I tried both GRM (Graded Response Model) and GPCM (Generalized Partial Credit Model), also both models unconstricted and with constricted “a” parameter for all 8 items, since they should all measure the same construct and there should be one latent variable estimate. According to the AIC, AICc, SABIC and BIC criteria, the best model with the least information lost was the constricted GRM model, which I will be using for all four scales (Table 4).

Table 4 - Assessing best polytomous model

	AIC	AICc	SABIC	BIC
GRM unconstricted	22691.94	22694.28	22748.86	22853.67
GRM constricted	22648.20	22651.65	22717.19	22844.23
GPCM unconstricted	22777.50	22779.84	22834.41	22939.22
GPCM constricted	22738.18	22741.63	22807.17	22934.21

With that said, in Figure 7 are the plotted item trace lines, in Figure 8 can be found the item information lines and in Figure 9, the test information line is plotted. In Table 5, all parameter estimates can be found. From the plot we can see that all items function normally – the higher the estimated ability (or, in this case, preference of the AG humor style), the higher probability of choosing a positive answer for each statement. For example, most information at higher ability

estimates is provided by the item Q27: *"If I don't like someone, I often use humor or teasing to put them down*" (Figure 8), which is also the item with least difficulty. This makes sense; only the people with the most prevalent aggressive humor style (and therefore highest estimated latent trait) will agree with this statement, as this is a socially undesirable behavior. Therefore the item well captures individuals with heavy use of the aggressive humor style.

Figure 7 - Constricted GRM model for the AG scale

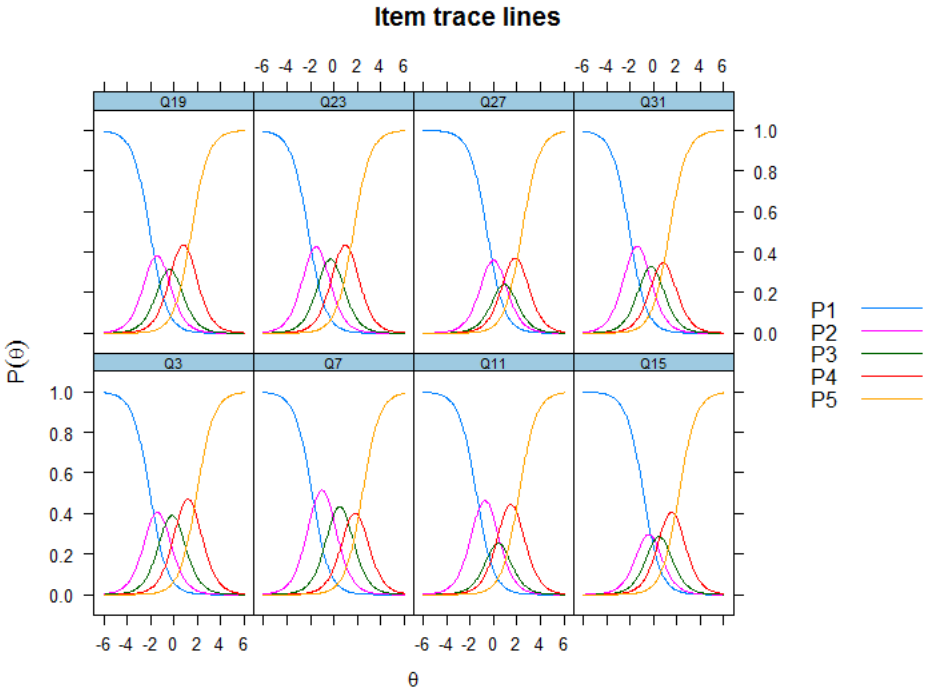


Figure 8 - Item information lines using the constricted GRM model for the AG scale

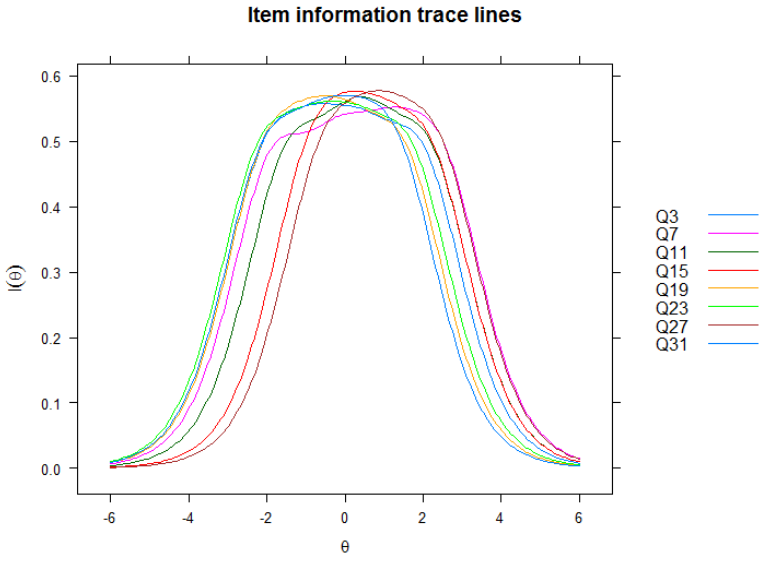


Figure 9 - Test information and SE using the constricted GRM model for the AG scale

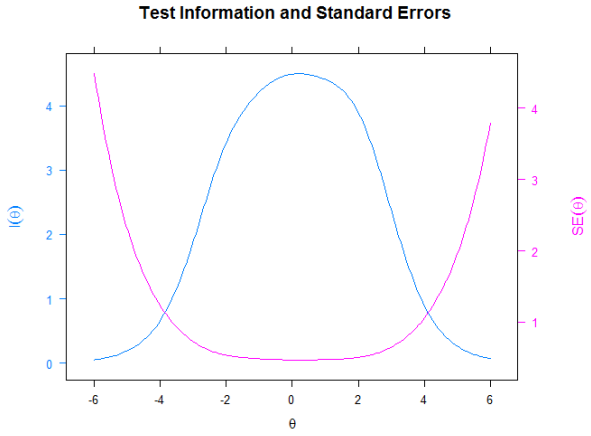


Table 5 - Constricted GRM parameter estimates, their standard errors & model fits for the AG scale

	a	SE(a)	b1	SE(b1)	b2	SE(b2)	b3	SE(b3)	b4	SE(b4)	S_X2 value	df	S_X2 p-value
Q3	1.353098	0.042501	2.7528575	0.1227059	1.0297170	0.0882963	-0.6168914	0.0855141	-2.660616	0.1196004	88.23905	70	0.0694210
Q7	1.353098	0.042501	2.5028621	0.1150437	0.2271619	0.0843067	-1.6199275	0.0957764	-3.322818	0.1443355	74.43794	70	0.3359582
Q11	1.353098	0.042501	2.0099694	0.1033244	0.0050191	0.0839332	-1.0302106	0.0885949	-2.940161	0.1291463	105.18463	71	0.0052365
Q15	1.353098	0.042501	1.1833343	0.0896490	-0.0387875	0.0838696	-1.2161635	0.0904354	-2.935429	0.1272456	79.62812	73	0.2784081
Q19	1.353098	0.042501	2.7520079	0.1231480	1.1432030	0.0897228	-0.1655792	0.0840971	-2.032040	0.1041942	62.75455	69	0.6884155
Q23	1.353098	0.042501	2.9468003	0.1301865	1.1317655	0.0896017	-0.3991716	0.0846750	-2.264564	0.1086379	105.46358	69	0.0031096
Q27	1.353098	0.042501	0.7698674	0.0866765	-0.7362581	0.0861506	-1.7184281	0.0980053	-3.273906	0.1436379	69.52386	70	0.4935850
Q31	1.353098	0.042501	2.8114988	0.1229472	0.9797247	0.0876398	-0.3788255	0.0844064	-1.820934	0.0989559	78.28103	71	0.2588770

IRT models with binarized data

Since a likert scale is well transformable into a binarized dataset, I decided to fit a dichotomous model for my data as well. When binarizing data, I coded all “4 = Often true” and “5 = Very often or always true” as “1” (agreement with the statement) and the rest “1 = Never or very rarely true, 2 = Rarely true, 3 = Sometimes true” as “0” (disagreement with the statement). I decided to code “3” as “0” since “sometimes” does not quite qualify as a prevalence of that humor style. Then I ran the same model comparison as in the previous section “Polytomous models” (see Table 6). While it is not clear which model fits best as in the previous section, I decided to use the 2PL model which fits best according to the SABIC and BIC criteria – mostly due to the simplicity of the model, number of parameters estimated and nature of the data (e.g. guessing or inattention are not as easily interpretable in personality tests).

Table 6 - Assessing best dichotomous model

	AIC	AICc	SABIC	BIC
1PL	7230.669	7230.852	7246.192	7274.776
2PL	7172.841	7173.398	7200.436	7251.252
3PL	7170.841	7172.081	7212.234	7288.459
4PL	7163.012	7165.212	7218.202	7319.835

Below you can find the 2PL IRT model for the AG scale (Figure 10), its information trace lines (Figure 11), test information and standard errors (Figure 12) and parameter estimates (Table 7). When we compare these dichotomous models to the polytomous models above, it is visible that the same items have similar characteristics – e.g. item "Q27" again provides most information about high latent trait estimates (Figure 11). Item Q19 on the other hand shows to have the largest probability of a positive answer for the low latent trait estimate respondents in the 2PL IRT model, corresponding with the traditional item analysis, where this item had the least difficulty. Overall, it looks like both models are similar in terms of item characteristics and both seem to be similarly interpretable.

Figure 10 - 2PL IRT model for the AG scale

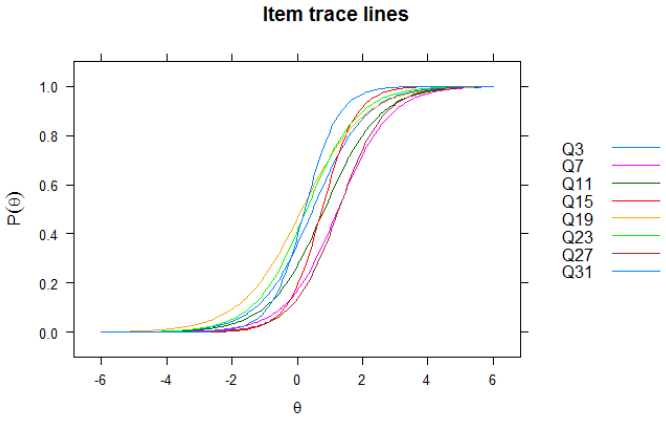


Figure 11 - Item information lines using the 2PL IRT model for the AG scale

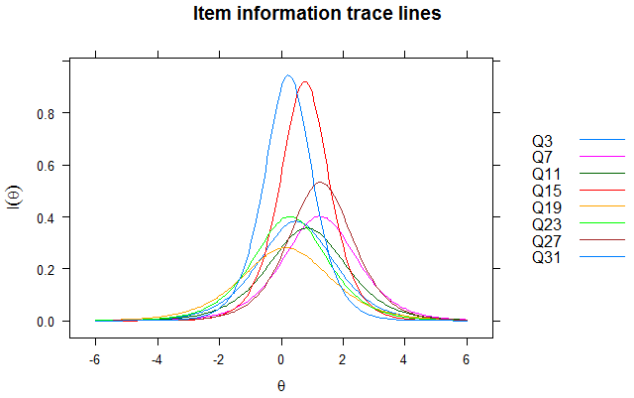


Figure 12 - Test information and SE using the 2PL IRT model for the AG scale

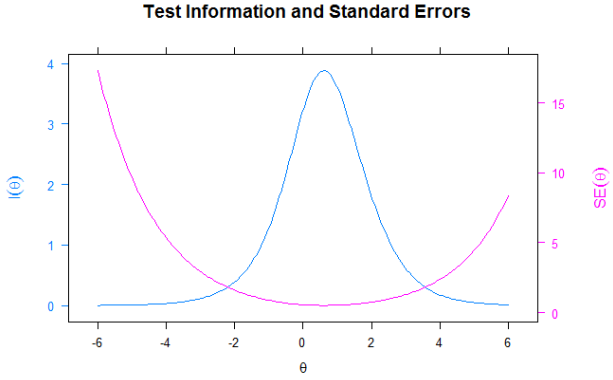
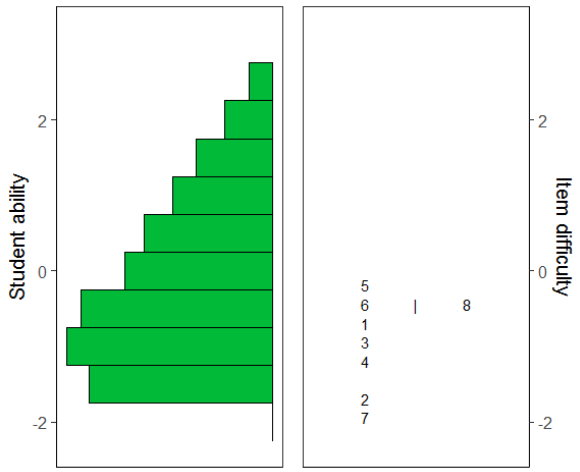


Table 7 – 2PL IRT parameter estimates, their standard errors & model fits for the AG scale

	a	SE(a)	b	SE(b)	S_X2 value	df	S_X2 p-value
Q3	1.238060	0.1277474	-0.6043211	0.0873302	6.375834	5	0.2713465
Q7	1.268160	0.1399313	-1.5947396	0.1153216	9.101067	5	0.1051002
Q11	1.193787	0.1271799	-1.0015267	0.0939932	8.749656	5	0.1194759
Q15	1.919436	0.2006393	-1.4631174	0.1377925	7.776932	5	0.1689654
Q19	1.060136	0.1141115	-0.1613891	0.0785102	2.978264	5	0.7033370
Q23	1.266951	0.1297699	-0.3697077	0.0851985	8.691251	5	0.1220311
Q27	1.460851	0.1597912	-1.8652628	0.1351938	1.342464	5	0.9305021
Q31	1.945204	0.2003904	-0.4562208	0.1067882	2.291819	5	0.8074677

In Figure 13 below, you can see a fitted Wright map displaying ability estimates on the left and item difficulties on the right (for the AG scale, for other scales, see Supplement R code). The items are marked in numerical order and shows difficulty for *binarized* items, though they correspond with the graph in the "Traditional item analysis" section. This plot also suggests that the AG humor style is not as prevalent due to lower average ability estimates, which corresponds with all previous analyses.

Figure 13 - Wright map for the AG scale



Differential Functioning Items (DIF)

The original dataset also includes gender and age variables for each participant. Therefore, I decided to look at items that function differently for men and women. To determine these items, methods used are the Lord and Raju IRT methods. Delta plot, logistic regression and other simpler models were not used because computation ability is not an issue and I want to be able to detect non-uniform DIF as well as uniform DIF. Before the analyses, I manipulated the data in a way that would leave me with only two groups – I removed 8 participants that selected "Other" gender and participants that selected no gender. Both methods (Raju and Lord) yielded the same results (marked the same items as DIF for each scale), therefore I will only show results from the Lord method for the

two scales that had DIF items (AG, AF) – rest can be found in Supplement materials. Scales SE and SD had no DIF items detected by either method.

In Figure 13 you can see the item that was detected as DIF in the AG scale – fifth analyzed item, that is Q19: "*Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation*." If we look at the item curve, it is interesting that it's *women* who more often answer positively to this question with high estimated preference for the AG humor style, but women who do not score high on this scale are much less likely than men to agree with this statement. This may reflect reality to the point that men more often tell jokes that may be inappropriate without the intention to hurt others.

Figure 14 - DIF items detected by the Lord IRT method for AG scale

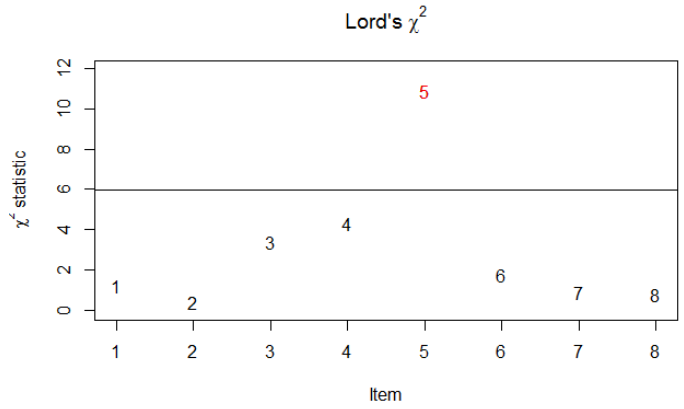
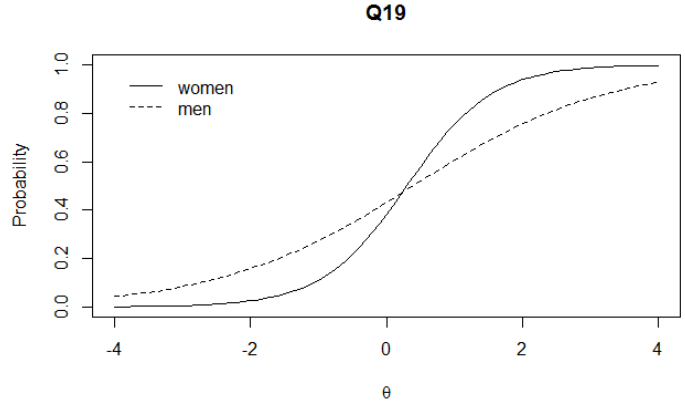


Figure 15 - ICC for item Q19 for both genders



The scale AF is interesting because there were two DIF items detected by both methods. In Figure 15, 16 and 17, you can see the detection test and ICCs, respectively. It is important to note that item Q5 ("*I don't have to work very hard at making other people laugh—I seem to be a naturally humorous person*") was barely significant ($p < .05$). That said, the result corresponds with the underlying idea that men often think they're funnier than they are. The item Q9 ("*I rarely make other people laugh by telling funny stories about myself*") also shows similar pattern, though I would assume just by observation in real life that it should be the opposite. It appears that men of higher preference for the AF humor style are less likely to see themselves as being funny to their surroundings, unlike women. With that said, I do not think that these items are biased in a way that would qualify them for refinement or removal. In fact, women and men *are* different and a

personality test will always reflect that (if it aims to measure personality traits that differ in any way based on gender). Thus, I think that this is a natural occurrence of different behavior in humor and it does not impair the quality of the questionnaire. It is also important to note that all three detected DIF items were non-uniform, meaning I used the correct methods to detect the differences that would not be detected by other, simpler methods.

Figure 16 - DIF items detected by the Lord IRT method for AF scale

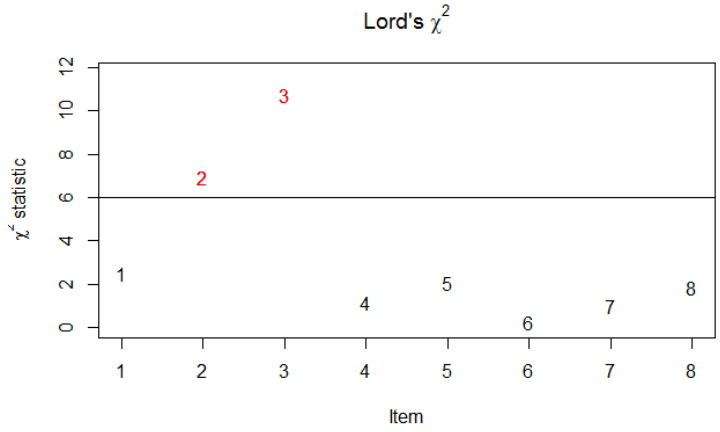


Figure 17 - ICC for item Q5 for both genders

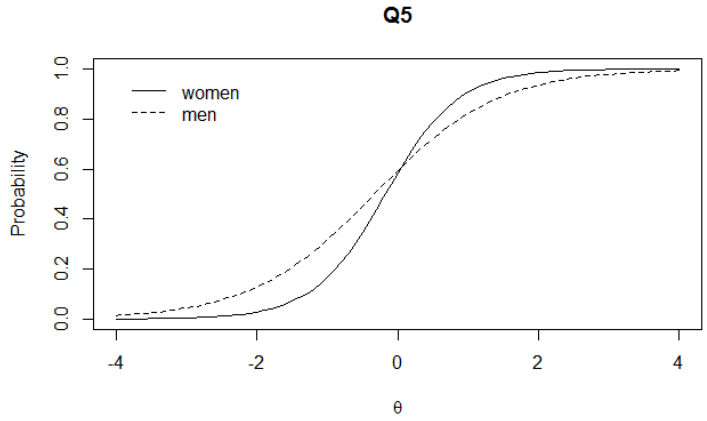
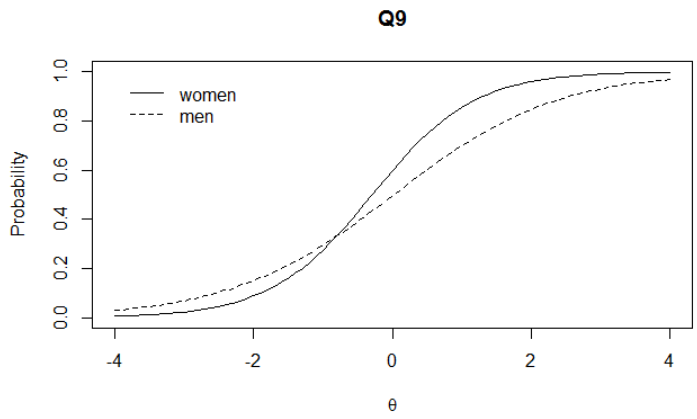


Figure 18 - ICC for item Q9 for both genders



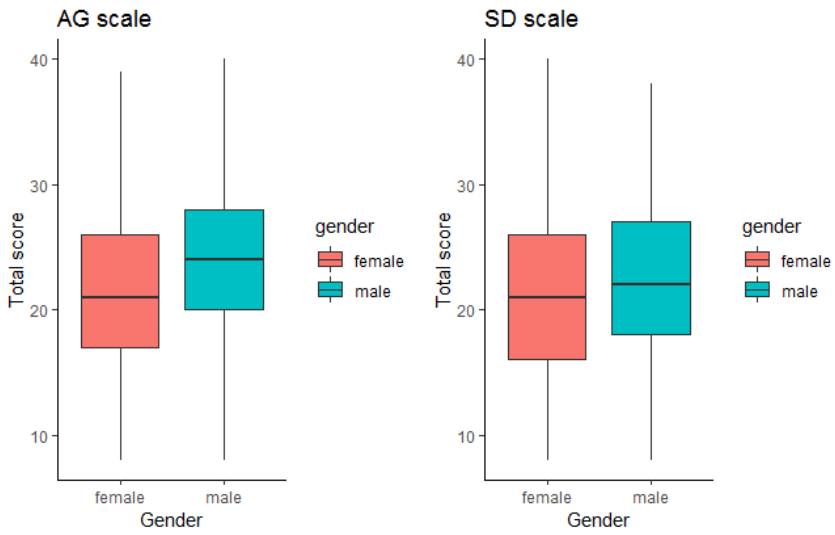
Factor analysis

I wanted to replicate an analysis that was done by Martin, et. al (2003) in the original paper – the Confirmatory Factor Analysis. Since the data is supposed to be split into four factors that measure different latent traits (styles of humor), I fit a CFA model. The model was on the border of an acceptable fit with RMSEA = .06 and SRMR = .067. However, according to the CFI = .84, the fit is not satisfactory (Bentler & Hu, 2009). It turns out that I did not manage to replicate the very good fit from the original article (RMSEA = .048), though my model seems to be acceptable under two out of three measures. Because of this outcome, I also fitted an EFA (Exploratory Factor Analysis) to see factor loadings of each item and it turns out all four scales work quite well – each item loaded into the factor it should and there were no overlaps, though some items showed to be worse than others. This is similar to the findings of Ruch & Heintz (2016), who used the German version of the questionnaire and obtained similar results in their PCA and CFA.

Further analyses of differences between participant groups

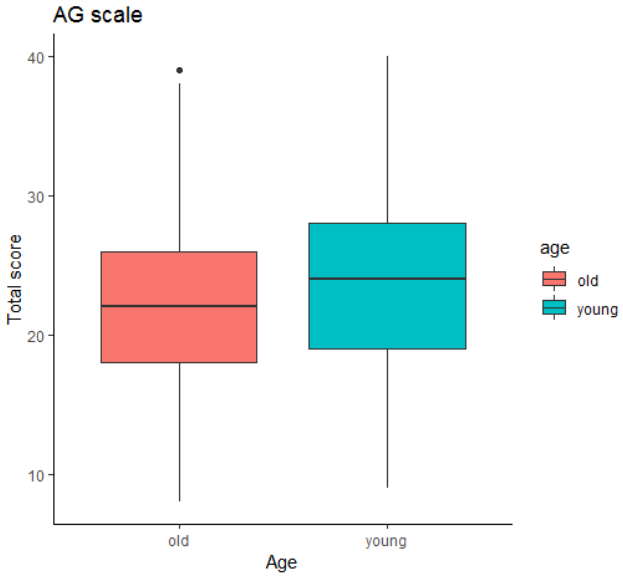
Apart from the analysis of the items, I wanted to look at differences between groups in my sample. As per Martin, et. al, men use the AG and the SD humor styles more than women ($p < .01$). To determine whether there is a significant difference in presented sample as well, I used Welch's two-sided t-test for all four scales. As expected, the AF and SE scales had no significant difference between total scores. On the other hand, men scored higher on average on the AG scale ($p < .001$) and on the SD scale ($p < .01$), which follows the original paper. For plotted results, see Figure 18.

Figure 19 - Differences between men and women on the AG & SD scales



Furthermore, I split the sample into two age groups – "old" (ages above 25) and "young" (ages below 25"). I chose the cut-off of 25 years per the original study; besides, mean age is 26. After the split, the "old" sample consists of 419 observations and "young" of 570 observations. Following the original paper, younger people have significantly higher scores on the AG scale ($p < .001$). On all other three scales, no significant differences were found.

Figure 20 - Differences between older and younger group of respondents on the AG scale



Discussion & further recommendations

In this analysis, I tried to assess the HSQ with more precise measures than the authors in the original article (Martin et al., 2003). I am confident in this tool as most analyses replicated successfully – the sample shows preference of certain scales by certain groups corresponding with the initial paper. The IRT models, both polytomous and dichotomous, showed good functioning of the items and the scales seem consistent in terms or reliability and validity. However, it is important to note that the CFA model did not fit as well as was expected, even though neither item loaded onto a different factor. Correlations between scales also did not show as satisfying results as in the original study; the AF and SE scales correlated well as expected, but there should be lower correlations between the positive and negative scales (AF and AG, SE and SD, etc.).

Unfortunately, I cannot compare the distribution of frequency of answers to the original paper, because that data is not available. That said, the mean total score is higher for the AF scale in the original study as well, which may imply that there is, in fact, a preference for this humor style. That could mean it is not an issue with the measure, rather than a good indicator of discrimination between healthy and unhealthy humor styles. The same argument shall apply to the DIF items; in personality tests, it is hard to determine whether gender differences in answers imply a poorly constructed tool, or whether it reflects actual gender differences. To further explore this topic, I suggest collecting data from a widely culturally different sample (African tribes, etc.) to see whether these differences are culture-laden or innate (eastern cultures seem to work similarly, see for example Chen & Martin, 2007)

With that said, I think there are a few tweaks that would only help the questionnaire. For example, the item Q19: *"Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation"* showed to be inconsistent with the scale with the highest difficulty but low discrimination in comparison. This item has also shown to provide most information for low estimated latent traits (AG style preference) and has been detected as DIF. Taking the wording into account, I suggest rewording this item. It needs to reflect the *intention* to hurt others, as per the definition of the AG style, such as: "*Sometimes I think of something that is so*

funny that I can't stop myself from saying it, even if I know it will probably hurt others." I also suggest tweaking the AF scale a little to reflect more accurately the preference for this humor style – while it probably still will be prevalent, the items should not be as universally "true". For example, item Q21: ("*I enjoy making people laugh*") has shown very high percentage of positive answers, even by those who do not prefer the AF style. How the item is written suggests it could include all other scales – even people with the preference for AG humor style *enjoy making people laugh*, but in an aggressive way. I suggest something along the lines of: "*I often use innocent jokes to put others at ease."* – that way, this item more specifically addresses the AF scale. I also find the item Q6: "*Even when I'm by myself, I'm often amused by the absurdities of life*" to be problematic, due to the very high difficulty, low discrimination and item information trace line suggesting this item provides most information on low latent trait estimate participants (see Supplement), implying total score on this item may correlate weakly with score on this item (Cronbach's alpha =). It hits the same issue as the items mentioned above – it may capture overall tendency of humans to laugh at the absurdities of life, which is consistent with the SE scale, but seems way too general. I suggest rewording along the lines of: "*Even when I'm by myself, I often find things to joke and laugh about alone*".

The limitations of my analysis are clearly due to the sample – the data was collected anonymously via the openpsychometrics.org website and anyone could enter. While there is little reason to "fake" answers, it is not a representative sample and we should not generalize the results. In the future, I highly suggest analyzing a sample from more cultures and possibly even create a Czech mutation of the HSQ, which may provide more insight into the Czech population.

References

Baughman, H. M., Giammarco, E. A., Veselka, L., Schermer, J. A., Martin, N. G., Lynskey, M., & Vernon, P. A. (2012). A Behavioral Genetic Study of Humor Styles in an Australian Sample. Twin Research And Human Genetics, 15(05), 663-667. http://doi.org/10.1017/thg.2012.23

Chen, G. -H., & Martin, R. A. (2007). A comparison of humor styles, coping humor, and mental health between Chinese and Canadian university students. Humor – International Journal Of Humor Research, 20(3). http://doi.org/10.1515/HUMOR.2007.011

Heintz, S., & Ruch, W. (2019). From four to nine styles: An update on individual differences in humor. Personality And Individual Differences, 2019(141), 7-12. http://doi.org/doi.org/10.1016/j.paid.2018.12.008

Ruch, W., & Heintz, S. (2016). The German version of the Humor Styles Questionnaire: Psychometric properties and overlap with other styles of humor. Europe'S Journal Of Psychology, 12(3), 434-455. http://doi.org/10.5964/ejop.v12i3.1116

Hu, Li-tze, & Bentler, P. (2009). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 2009(6), 1-55. http://doi.org/10.1080/10705519909540118

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. Journal Of Research In Personality, 37(1), 48-75. http://doi.org/10.1016/S0092-6566(02)00534-2

Supplement materials

Full list of items

Scale	Statement
AF	Q1. I usually don't laugh or joke around much with other people.
SE	Q2. If I am feeling depressed, I can usually cheer myself up with humor.
AG	Q3. If someone makes a mistake, I will often tease them about it.
SD	Q4. I let people laugh at me or make fun at my expense more than I should.
AF	Q5. I don't have to work very hard at making other people laugh—I seem to be a naturally humorous person.
SE	Q6. Even when I'm by myself, I'm often amused by the absurdities of life.
AG	Q7. People are never offended or hurt by my sense of humor.
SD	Q8. I will often get carried away in putting myself down if it makes my family or friends laugh.
AF	Q9. I rarely make other people laugh by telling funny stories about myself.
SE	Q10. If I am feeling upset or unhappy I usually try to think of something funny about the situation to make myself feel better.
AG	Q11. When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it.
SD	Q12. I often try to make people like or accept me more by saying something funny about my own weaknesses, blunders, or faults.

AF	Q13. I laugh and joke a lot with my closest friends.
SE	Q14. My humorous outlook on life keeps me from getting overly upset or depressed about things.
AG	Q15. I do not like it when people use humor as a way of criticizing or putting someone down.
SD	Q16. I don't often say funny things to put myself down.
AF	Q17. I usually don't like to tell jokes or amuse people.
SE	Q18. If I'm by myself and I'm feeling unhappy, I make an effort to think of something funny to cheer myself up.
AG	Q19. Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation.
SD	Q20. I often go overboard in putting myself down when I am making jokes or trying to be funny.
AF	Q21. I enjoy making people laugh.
SE	Q22. If I am feeling sad or upset, I usually lose my sense of humor.
AG	Q23. I never participate in laughing at others even if all my friends are doing it.
SD	Q24. When I am with friends or family, I often seem to be the one that other people make fun of or joke about.
AF	Q25. I don't often joke around with my friends.
SE	Q26. It is my experience that thinking about some amusing aspect of a situation is often a very effective way of coping with problems.
AG	Q27. If I don't like someone, I often use humor or teasing to put them down.
SD	Q28. If I am having problems or feeling unhappy, I often cover it up by joking around, so that even my closest friends don't know how I really feel.
AF	Q29. I usually can't think of witty things to say when I'm with other people.
SE	Q30. I don't need to be with other people to feel amused – I can usually find things to laugh about even when I'm by myself.
AG	Q31. Even if something is really funny to me, I will not laugh or joke about it if someone will be offended.
SD	Q32. Letting others laugh at me is my way of keeping my friends and family in good spirits.

Traditional item analysis – all scales

	Scale	Difficulty	SD	Discrimination ULI	Discrimination RIT	Discrimination RIR	Alpha Drop	Customized Discrimination
Q1	AF	3.97	1.06	1.64	0.72	0.61	0.82	0.41
Q5	AF	3.62	1.04	1.59	0.71	0.60	0.82	0.40
Q9	AF	3.40	1.21	1.69	0.63	0.47	0.84	0.42
Q13	AF	4.46	0.84	1.11	0.66	0.56	0.82	0.28
Q17	AF	4.06	1.10	1.84	0.79	0.69	0.80	0.46
Q21	AF	4.40	0.85	1.24	0.70	0.61	0.82	0.31
Q25	AF	4.45	0.84	1.20	0.70	0.61	0.82	0.30
Q29	AF	3.67	1.18	1.74	0.65	0.50	0.83	0.44

	Scale	Difficulty	SD	Discrimination ULI	Discrimination RIT	Discrimination RIR	Alpha Drop	Customized Discrimination
Q2	SE	3.35	1.09	1.65	0.69	0.57	0.80	0.41
Q6	SE	4.18	0.94	1.16	0.57	0.45	0.81	0.29
Q10	SE	2.88	1.18	1.97	0.75	0.64	0.79	0.49
Q14	SE	3.30	1.24	2.09	0.76	0.64	0.79	0.52
Q18	SE	2.76	1.17	1.87	0.73	0.62	0.79	0.47
Q22	SE	2.96	1.19	1.43	0.54	0.37	0.83	0.36
Q26	SE	3.57	1.13	1.79	0.73	0.62	0.79	0.45
Q30	SE	3.99	1.05	1.27	0.56	0.42	0.82	0.32

	Scale	Difficulty	SD	Discrimination ULI	Discrimination RIT	Discrimination RIR	Alpha Drop	Customized Discrimination
Q3	AG	3.08	1.16	1.57	0.64	0.51	0.77	0.39
Q7	AG	2.71	1.08	1.47	0.61	0.49	0.77	0.37
Q11	AG	2.73	1.24	1.68	0.61	0.46	0.77	0.42
Q15	AG	2.57	1.34	2.14	0.72	0.58	0.75	0.53
Q19	AG	3.24	1.23	1.55	0.59	0.44	0.78	0.39
Q23	AG	3.20	1.18	1.54	0.59	0.44	0.77	0.39
Q27	AG	2.26	1.26	1.69	0.62	0.47	0.77	0.42
Q31	AG	3.21	1.28	2.09	0.71	0.59	0.75	0.52

	Scale	Difficulty	SD	Discrimination ULI	Discrimination RIT	Discrimination RIR	Alpha Drop	Customized Discrimination
Q4	SD	2.82	1.16	1.74	0.70	0.58	0.79	0.43
Q8	SD	2.54	1.19	2.07	0.77	0.67	0.78	0.52
Q12	SD	2.95	1.21	1.91	0.70	0.58	0.79	0.48
Q16	SD	2.86	1.19	1.72	0.65	0.51	0.80	0.43
Q20	SD	2.08	1.09	1.80	0.75	0.65	0.78	0.45
Q24	SD	2.44	1.12	1.29	0.55	0.41	0.82	0.32
Q28	SD	3.22	1.30	1.41	0.49	0.31	0.83	0.35
Q32	SD	2.84	1.23	2.02	0.74	0.62	0.79	0.50

Item information trace lines for the SE scale