TwinLife Working Paper Series

No. 02, May 2018

Item response theory analysis of the cognitive ability test in TwinLife

by Sarah Carroll[1, 2] & Eric Turkheimer[1]

[1] Department of Psychology, University of Virginia

[2] Email corresponding author: slc4fv@virginia.edu

Sarah Carroll & Eric Turkheimer

Item response theory analysis of the cognitive ability test in TwinLife

TwinLife Working Paper Series No. 02
Project TwinLife “Genetic and social causes of life chances”
Bielefeld, May 2018

TwinLife Working Paper Series No. 02
General Editors: Martin Diewald, Rainer Riemann and Frank M. Spinath
ISSN 2512-4048

TwinLife is funded by the German Research Foundation (DFG).

TwinLife Working Papers are refereed scholarly papers. Submissions are reviewed by the general editors before a final decision on publication is made.

The Working Paper Series is a forum for presenting works in progress. Readers should communicate comments on the manuscript directly to the author(s).

The papers can be downloaded from the project website:
http://www.twin-life.de/en/twinlife-working-paper-series

TwinLife “Genetic and social causes of life chances”
University of Bielefeld
Faculty of Sociology
PO Box 100131
D-33501 Bielefeld
Germany

Phone: +49 (0)521 106-4309
Email: martin.diewald@uni-bielefeld.de
Web: http://www.twin-life.de/en

Universität Bielefeld

UNIVERSITÄT DES SAARLANDES

Item response theory analysis of the cognitive ability test in TwinLife

Sarah Carroll[1] and Eric Turkheimer[1]

[1] Department of Psychology, University of Virginia, Box 400400, Charlottesville, VA 22904, USA

TwinLife, an ongoing German study of twins and their families, investigates cognitive performance as one factor among many that contribute to the development of social inequality. Participants completed the CFT 20-R, a nonverbal intelligence assessment. The current analysis applied a two-parameter logistic item response theory model using Mplus software to subtest results from twin pairs in the three oldest birth cohorts, ranging in age from 10 to 25 years old. The findings indicated that the 2PL model fit the data considerably better than the one-parameter logistic model did for all four of the CFT 20-R subtests used in TwinLife. Results from the 2PL model, including item and person parameters and test information, are discussed. In addition, the items were assessed for measurement invariance across age cohort and gender. Fit statistics reveal little difference in item function according to these demographic factors, meaning that the CFT 20-R may be valid in heterogeneous samples.

Keywords: Item response theory; measurement invariance; cognitive ability; behavior genetics

Correspondence concerning this article should be addressed to Sarah Carroll, Department of Psychology, University of Virginia, USA, E-mail: slc4fv@virginia.edu

Introduction

Initiated in 2014, TwinLife is an ongoing behavior genetic study of social inequality in Germany. It investigates factors such as intelligence, educational attainment, and physical and mental health, which are expected to contribute to differing life outcomes among participants over a nine-year period (Diewald et al., 2016; Hahn et al., 2016).

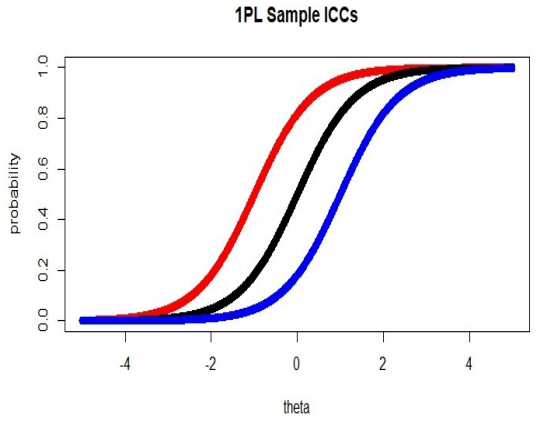
The current study focuses on differences in intellectual ability. Participants from the three oldest birth cohorts completed four subtests from the CFT 20-R, a timed, nonverbal intelligence assessment (Weiß, 2006). The Figural Reasoning, Figural Classification, and Matrices subtests each have 15 five-option multiple choice questions; the Reasoning subtest has 11 (Gottschling, 2017). The purpose of this analysis was to use item response theory methods to assess performance on the subtests.

Item response theory is a paradigm that relates an individual's trait level (θ) to his or her performance on a series of items, while accounting for item characteristics. Two item characteristics, item difficulty (β) and item discrimination (α), are considered in the IRT models we estimate here. Item discrimination indicates the extent to which performance on a given item relates to one's trait level; the higher the discrimination, the more accurately the item assesses ability. In IRT models, trait level and item difficulty are measured on the same scale; if an individual's trait level is equal to a given item's difficulty, then he or she has a 50% chance of answering the item correctly. This relationship is depicted in the item characteristic curve (ICC), a graph of the probability of endorsing an item given one's trait level and the item's properties (de Ayala, 2009).

For data that are coded dichotomously as correct or incorrect, either a one-parameter or two-parameter logistic model may be used. In the 1PL model, the items are free to vary in their difficulty but are assigned the same discrimination value; the 2PL model allows the items to vary by both difficulty and discrimination (de Ayala, 2009). One-parameter models offer significant advantages when they fit the data. The 1PL and 2PL models are shown in equations 1 and 2, respectively, and sample ICCs are shown in Figures 1 and 2. The sample ICCs in Figure 1, generated using the 1PL model, are parallel, indicating that they have the same discrimination, or slope, while the ICCs for the 2PL model in Figure 2 vary in their discrimination values. The location of the inflection point of each curve along the x-axis indicates item difficulty; the further the curve is shifted to the right, the more difficult the item is.

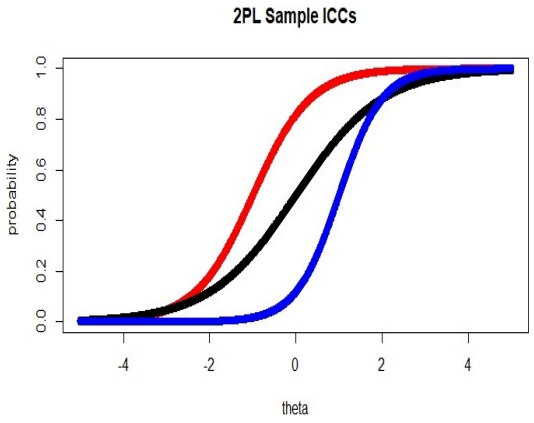
$$P_{ij}(\theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_j)} \quad (1)$$

Figure 1: *Sample ICCs for 1PL Model.*



$$P_{ij}(\theta_j, b_i, a_i) = \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_j)]} \quad (2)$$

Figure 2: *Sample ICCs for 2PL Model.*



Analyses include a model fit comparison of the 1PL and 2PL models for each subtest, followed by a discussion of item and person parameters. Items were assessed for measurement invariance by age and gender. Measurement invariance is observed when items relate to the latent trait consistently across groups of participants, after controlling for intergroup differences in average ability. Results confirming consistency of item functioning indicate that they measure the same construct across groups (Reise, Widaman, & Pugh, 1993). Differential item functioning is often detected when groups of participants perform differently on test items (Steinberg & Thissen, 2006). We evaluated item function across age and gender due to age-related differences in cognitive performance and differences between men and women in performance on spatial reasoning (Linn & Petersen, 1986), an ability that may be relevant given the nonverbal, figure-based nature of the four subtests.

Method

Participants

Participants were 6148 German citizens (3074 twin pairs) aged 10 years and older who completed the cognitive test battery on the computer in 2014. 1441 of the pairs were monozygotic and 1633 were dizygotic. All pairs were same-sex. Although the TwinLife sample includes four birth cohorts, Cohort 1 was excluded from these analyses because participants completed the CFT 1-R, a child version of the CFT 20-R (see Gottschling, 2017). There were 1036, 1058, and 980 twin pairs in Cohorts 2, 3, and 4, respectively. At the time of testing, participants in Cohort 2 ranged in age from 10 to 12 years, with a mean age of 11.00 years, while those in Cohort 3 ranged from 16 to 18 years with a mean of 17.00. Members of Cohort 4 had a mean age of 23.04 years, with a range of 21 to 25 years. The majority of participants in each cohort were women. Cohorts 2, 3, and 4 were 51.93%, 57.28%, and 58.16% female, respectively.

Data Analysis

Analyses were carried out in Mplus 7.4 (Muthén & Muthén, 1998-2015). The models accounted for the correlation within twin pairs but did not employ a traditional twin design. The initial dataset containing all participants' responses to the subtest was reduced to exclude members of Cohort 1 and non-twin participants. The first step of the analysis was a fit comparison of the 1PL and 2PL models for each of the four subtests. We assessed model fit using two different statistics: a chi-square difference test which, when significant, indicates that the more constrained model fits the data significantly worse than the less constrained model, and the root mean square error of approximation (RMSEA). RMSEA values below 0.06 are considered to indicate good fit. When the two tests gave conflicting results, we relied on the RMSEA value because it is less sensitive to sample size than the chi-square value is. In a sample as large as the one used in these analyses, a negligible difference in fit could yield a significant chi-square value (Hooper et al., 2008).

Next, we tested for measurement invariance by cohort in each subtest by comparing model fit when parameters were constrained to be equal across age groups versus when they were free to vary. We tested for measurement invariance by gender by comparing model fit when parameters were constrained to be equal for male and female participants and when they were free to vary. Model fit was assessed using the same statistics described above. Because the RMSEA fit index indicated little difference in the models for either age or gender in any of the subtests, subsequent analyses collapsed across these groups.

Using the 2PL model, we estimated item discrimination and difficulty with the weighted least squares mean- and variance-adjusted (WLSMV) estimator (Muthén & Muthén, 1998-2015). We generated an ability (θ) estimate for each participant using the maximum likelihood estimator. In each subtest, we constrained the mean ability score to be equal to 0 and the variance to 1 for members of Cohort 2. Mean ability and variance were free to vary in cohorts

3 and 4, allowing us to identify age-related differences in average ability and spread of scores. The amount of information an item provides about ability level increases with its discrimination and is depicted in the item information curve, a graph of the relationship between person ability and item information. Information curves for all items on a test are summed together to create a test information curve, the location of which along the x-axis indicates the theta level where the test is most informative (de Ayala, 2009). We include a test information curve in our analyses of each subtest.

Results & Discussion

Figural Reasoning

Model Comparison: 1PL v. 2PL

We applied the 1PL and 2PL models to the data and performed a chi-square test to compare the fit of the nested models. The chi-square value, 847.992, was significant at $p<0.05$, indicating that the 2PL fits the data better than the 1PL.

The RMSEA values for the 2PL and 1PL models were 0.023 (0.020-0.026) and 0.052 (0.050-0.054), respectively. The parenthetical numbers following the RMSEA values represent the 90% confidence intervals. Both fit statistics, chi-square and RMSEA, indicate that the 2PL is the better fitting model. Subsequent analyses of the Figural Reasoning subtest use the 2PL model.

Measurement Invariance

For the two demographic factors of interest, age and gender, we compared model fit when parameters were free to vary by age group and gender and when they were constrained to be equal. Although a chi-square test of nested models indicated that the model in which parameters were free to vary by age fit significantly better than the constrained model, with a chi-square value of 328.340, the RMSEA value of 0.031 (0.029-0.033) for the latter indicated that the constrained model did not fit the data poorly.

The same pattern emerged for gender. The constrained model fit the data significantly worse, based on the chi-square value of 50.415, but its RMSEA of 0.024 (0.021-0.026) meant that fit was not poor, indicating that the items functioned similarly in men and women.

Item Parameters

Parameters were estimated using the 2PL model, collapsing across age and gender. Item difficulties for all subtests are shown at the end of the paper in Table 2. For the Figural Reasoning subtest, difficulties ranged from -2.298 to 1.020, with a mean of -0.895. Item 1 was the easiest, and Item 15 was the most difficult. Response patterns for items 13, 14, and 15, the most difficult on the test, differed by cohort. More than 50% of respondents in cohorts 3 and 4 answered these items correctly, while fewer than 35% in Cohort 2 did so. Response rates were similar across the three cohorts, with more than 85% of participants answering each question.

Item discriminations for all four subtests are shown in Table 3. All values were positive, meaning that a correct response to any item was associated with a higher score on the latent trait (de Ayala, 2009). For the Figural Reasoning subtest, Item 3 was the most discriminating, with an α of 0.814, while Item 8 was the least discriminating, with an α of 0.297. The mean was 0.529. Easier items, on average, discriminated better than harder ones did. The correlation between item difficulty and item discrimination values was -0.451 across cohorts.

Person Parameter

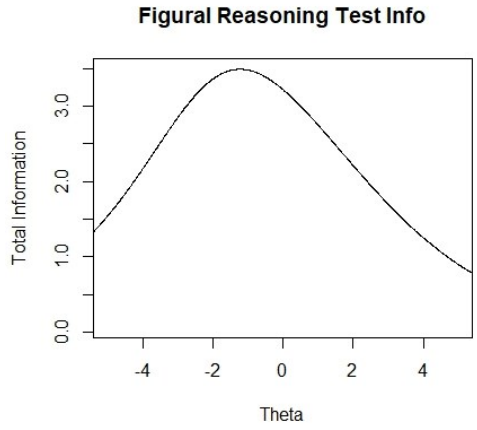
Person ability scores for each subtest are shown in Table 4. Mean ability was fixed at 0 in Cohort 2, with a variance of 1, for all subtests; the mean and variance were free to vary in cohorts 3 and 4. The mean score in Cohort 3 on the Figural Reasoning test was 0.960 (0.048), with a variance of 1.059 (0.072), and the mean in Cohort 4 was 0.878 (0.049), with a variance of 1.209 (0.072). Mean ability increased from Cohort 2 to Cohort 3 but declined slightly in Cohort 4, indicating a larger difference in academic ability between 11 and 17-year-olds than between 17 and 23-year-olds. The variance was slightly

higher in Cohort 4, indicating a wider range of ability levels among older participants on this subtest.

Test Information Curve

Figure 3 contains the test information curve, representing the total information about ability level provided by the test. This subtest is most informative for participants whose theta falls near -1.

Figure 3: *Test information curve for Figural Reasoning.*



Figural Classification

Model Comparison: 1PL v. 2PL

Based on both fit statistics used in these analyses, the 2PL model fit the data better than the 1PL. A chi-square test of nested models was significant at $p<0.05$, with a value of 874.827. The RMSEA values for the 2PL and 1PL models were 0.029 (0.026-0.031) and 0.055 (0.052-0.057), respectively. Subsequent analyses use the 2PL model.

Measurement Invariance

A chi-square test of nested models indicated that the model in which parameters were constrained to be equal across age cohorts fit significantly worse than the unconstrained model, with a value of 318.490. Based on the RMSEA value of 0.034 (0.032-0.036) for the constrained model, however, fit was not poor, indicating that the items functioned consistently in different age groups. There was also little evidence of differential item function by gender; the model in which parameters were constrained to be equal for men and women had an RMSEA of 0.031 (0.029-0.033), despite a significant chi-square value of 123.886.

Item Parameters

Item parameters were estimated using the constrained 2PL model, in which items were assumed to function consistently across age and gender. Item difficulties ranged from -1.951 (Item 5) to 10.534 (Item 15), with a mean of 0.391. Item 15 did not discriminate well among participants and appeared to be too difficult for this sample; roughly 10% of respondents in each cohort provided the correct answer, a rate lower than chance since participants were choosing among five options. The large standard error for this item's difficulty, included in Table 2, indicates that the estimate was less precise than it was for the other items. Discriminations ranged from 0.128 (Item 15) to 0.793 (Item 1), with a mean of 0.444. The correlation between item discrimination and difficulty was -0.762.

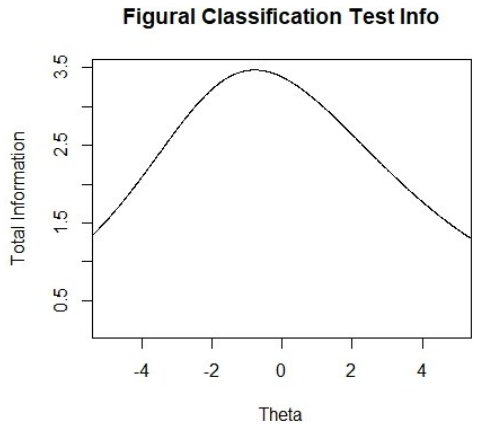
Person Parameter

Fixed at 0 in Cohort 2, the mean theta increased to 1.180 (0.053) in Cohort 3 and 1.094 (0.056) in Cohort 4. The variance also increased from 1 in Cohort 2 to 1.246 (0.077) and 1.466 (0.085) in cohorts 3 and 4, respectively, indicating a greater spread of scores among older participants.

Test Information Curve

Figure 4 contains the test information curve, which shows that the test is most informative at a theta between 0 and -1.

Figure 4: *Test information curve for Figural Classification.*



Matrices

Model Comparison: 1PL v. 2PL

The 2PL model fit the data significantly better than the 1PL did at $p<0.05$, with a chi-square of 1246.073. The RMSEA for the 2PL was 0.044 (0.041-0.046) and for the 1PL was 0.068 (0.066-0.070). Because the chi-square test and RMSEA values both indicate that the 2PL fits better than the 1PL, subsequent analyses of the Matrices subtest use the 2PL model.

Measurement Invariance

When parameters were free to vary across age cohorts, the model fit significantly better at $p<0.05$, according to a chi-square value of 238.662. The RMSEA, however, was 0.040 (0.038-0.042), indicating that the parameters could be constrained by age. Similarly, a chi-square test, with a value of 45.993, indicated that the model fit significantly worse when parameters were forced to be equal for male and female participants, although the RMSEA value of 0.039 (0.037-0.041) for the constrained model indicated little difference in item function by gender. Subsequent analyses of responses to the Matrices subtest collapse across age and gender.

Item Parameters

Item difficulties ranged from -2.186 to 1.099, with a mean of -0.589. Item 4 was the easiest, and Item 15 was the most difficult. The percentage of test-takers responding correctly to the most difficult items varied by cohort, with only 26.8% of respondents in Cohort 2 correctly answering Item 15 while roughly half answered correctly in the older cohorts. Item discriminations ranged from 0.370 to 0.881, with a mean of 0.627. Item 3 had the highest discrimination, and Item 9 had the lowest. The correlation between item difficulty and discrimination was -0.705.

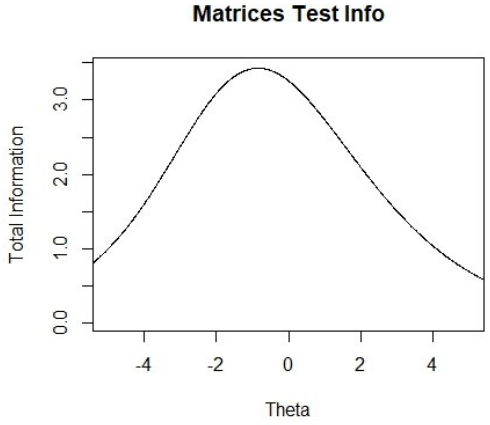
Person Parameter

Mean theta increased from 0 in Cohort 2 to 0.951 (0.045) in Cohort 3. It declined to 0.923 (0.046) in Cohort 4, consistent with the results from the other subtests. Fixed at 1 in Cohort 2, the variance increased slightly to 1.049 (0.048) in Cohort 3 and 1.046 (0.046) in Cohort 4, indicating a similar spread of scores across age groups.

Test Information Curve

Figure 5 contains the test information curve, which shows that the subtest is most informative for participants whose theta is between -1 and 0.

Figure 5: *Test information curve for Matrices.*



Reasoning

Model Comparison: 1PL v. 2PL

The 2PL model, with an RMSEA of 0.030 (0.026-0.033), fit the data better than the 1PL, which had an RMSEA of 0.076 (0.073-0.079). Additionally, a chi-square test of nested models yielded a significant result, 975.045, at $p<0.05$, indicating worse fit of the 1PL. Subsequent analyses use the 2PL model.

Measurement Invariance

When the 2PL model was constrained across age cohorts, it fit significantly worse than the unconstrained model, based on a significant chi-square value of 192.231. However, the RMSEA of 0.034 (0.031-0.037) for the constrained model indicated little difference in item function by age. Similarly, a test for measurement invariance by gender yielded a significant chi-square result, 56.566, but the RMSEA of 0.027 (0.024-0.030) for the constrained model meant that the items functioned consistently in men and women.

Item Parameters

Item difficulties, collapsed across age and gender, ranged from -1.402 (Item 1) to 2.881 (Item 11), with a mean of 0.430. Fewer than 10% of respondents in Cohort 2 answered Item 11 correctly, while roughly 25% from the two older cohorts responded correctly, a rate not far from

what would be expected if participants were guessing among the five options. Response rates differed by cohort; 77% of participants in Cohort 2 provided a response to Item 11, while 65% and 59% provided a response in cohorts 3 and 4, respectively. Because items left blank are coded as missing, not incorrect, this may indicate that older participants are better at gauging their own ability level.

Item discriminations ranged from 0.278 (Item 10) to 0.667 (Item 4), with a mean of 0.465. The correlation between item discriminations and difficulties was -0.197.

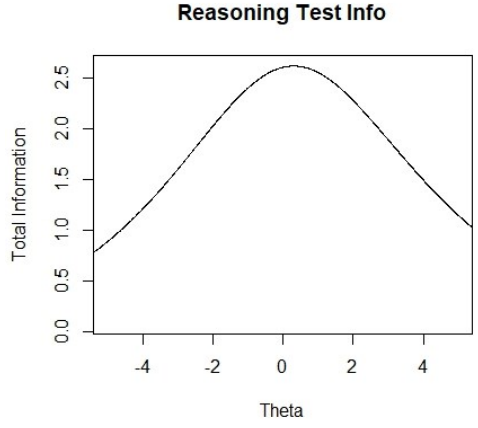
Person Parameter

Mean theta increased from 0 in Cohort 2 to 0.979 (0.057) in Cohort 3 and 1.119 (0.060) in Cohort 4. Unlike on the other subtests, where mean ability plateaued between cohorts 3 and 4, scores continued increasing with age. The variance was 1.640 (0.077) in Cohort 3 and 1.598 (0.077) in Cohort 4, indicating a larger spread of scores than in Cohort 2, where the variance was fixed to 1.

Test Information Curve

Figure 6 contains the test information curve, which peaks at a theta between 0 and 1. The Reasoning subtest, which contains fewer items than the other three subtests, also provides less total information, indicated by its location on the y-axis.

Figure 6: *Test information curve for Reasoning.*



Performance across subtests

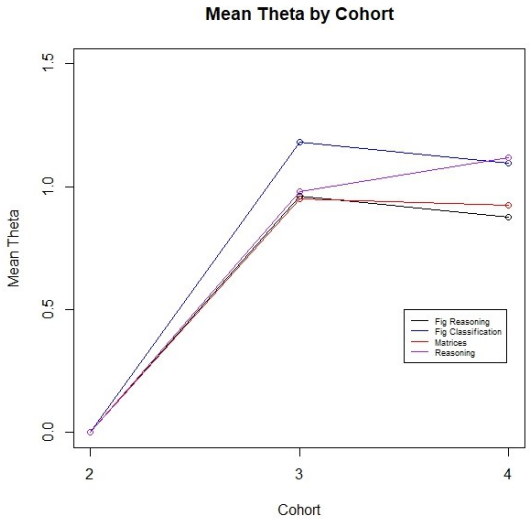
5984 participants had complete data for all four subtests. Participants' scores on the subtests, calculated as thetas, were moderately correlated, as shown in Table 1. Correlations were higher among scores on the first three subtests than for those on the Reasoning subtest.

Although the mean theta was fixed at 0 in Cohort 2 for each subtest, it showed a slightly different pattern of change with age across subtests, as shown in Figure 7. Mean theta increased from Cohort 2 to Cohort 3 in all subtests, while it declined slightly from Cohort 3 to Cohort 4 in the first three subtests. Only for the Reasoning test did the mean score continue to increase from Cohort 3 to Cohort 4.

Table 1: *Score correlations across subtests.*

	Figural Classification	Matrices	Reasoning
Figural Reasoning	0.619	0.647	0.528
Figural Classification		0.650	0.563
Matrices			0.556

Figure 7: Mean theta across cohorts in all subtests.



Conclusion

In the preceding analyses, the 2PL model was determined to best fit the data and was used to evaluate the items for measurement invariance according to two demographic factors: age and gender. The items were largely invariant, so parameters were estimated using a model that collapsed across these groups. Results were consistent across subtests, with mean ability increasing between cohorts 2 and 3 and easier items discriminating better than difficult ones. Despite the consistency in item function across age groups, the most difficult items on each subtest appeared to be too difficult to assess ability accurately in members of Cohort 2. Item 15 on the Figural Classification subtest and Item 11 on the Reasoning subtest may be too difficult for older participants as well. The majority of items on all four subtests appear to be appropriate for this sample, based on the range of scores in each cohort and the test information curves, which indicate that each subtest is most informative about participants whose ability levels fall between -1 and 1. In addition, the consistency in item function across age and gender means that the CFT 20-R may be appropriate for use in diverse samples such as TwinLife.

References

de Ayala, R. J. (2009). Methodology in the social sciences. The theory and practice of item response theory. New York.

Diewald, M., Riemann, R., Spinath, F. M., Gottschling, J., Hahn, E., Kornadt, A. E., … Peters, A.-L. (2016). TwinLife. GESIS Data Archive. https://doi.org/10.4232/1.12665.

Gottschling, J. (2017). TwinLife Technical Report Series, 02. Project TwinLife: Genetic and social causes of life chances (Universität Bielefeld / Universität des Saarlandes).

Hahn, E., Gottschling, J., Bleidorn, W., Kandler, C., Spengler, M., Kornadt, A. E., ... & Lang, V. (2016). What drives the development of social inequality over the life course? The German TwinLife Study. Twin Research and Human Genetics, 19(6), 659-672.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. Articles, 2.

Linn, M. C., & Petersen, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. The psychology of gender: Advances through meta-analysis, 67-101.

Muthén, L.K. and Muthén, B.O. (1998-2015). Mplus User's Guide. Seventh Edition.

Los Angeles, CA: Muthén & Muthén.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. Psychological Methods, 11(4), 402.

Weiß, R. (2006). CFT 20-R. Grundintelligenztest Skala 2. Manual. Göttingen: Hogrefe.

Appendix

Table 2: *Item difficulties for all subtests.*

Item Difficulties (S.E.)				
	Figural Reasoning	Figural Classification	Matrices	Reasoning
Item 1	-2.298 (.141)	-1.782 (.091)	-2.033 (.095)	-1.402 (.082)
Item 2	-1.885 (.092)	-1.364 (.074)	-1.933 (.082)	-.510 (.064)
Item 3	-1.848 (.086)	-1.491 (.079)	-1.907 (.080)	.879 (.046)
Item 4	-2.138 (.176)	-1.281 (.079)	-2.186 (.117)	.068 (.038)
Item 5	-2.239 (.141)	-1.951 (.134)	.139 (.045)	.304 (.043)
Item 6	-.981 (.054)	-.360 (.065)	-.775 (.046)	-.122 (.042)
Item 7	-1.257 (.076)	-.661 (.067)	-.541 (.037)	.646 (.059)
Item 8	-.462 (.087)	.296 (.052)	-.485 (.041)	-.492 (.071)
Item 9	-.714 (.072)	.415 (.053)	-.082 (.058)	1.212 (.062)
Item 10	-.131 (.052)	1.600 (.087)	-.169 (.052)	1.261 (.085)
Item 11	-1.326 (.103)	.052 (.051)	-.257 (.045)	2.881 (.138)
Item 12	-.531 (.057)	.361 (.051)	.459 (.044)	N/A
Item 13	.652 (.049)	-.144 (.057)	-.545 (.051)	N/A
Item 14	.709 (.044)	1.634 (.093)	.386 (.039)	N/A
Item 15	1.020 (.054)	10.534 (1.934)	1.099 (.051)	N/A

Table 3: *Item discriminations for all subtests.*

Item Discriminations (S.E.)				
	Figural Reasoning	Figural Classification	Matrices	Reasoning
Item 1	.719 (.033)	.793 (.026)	.793 (.024)	.466 (.016)
Item 2	.806 (.027)	.611 (.019)	.798 (.022)	.405 (.016)
Item 3	.814 (.027)	.636 (.021)	.881 (.023)	.625 (.016)
Item 4	.346 (.021)	.524 (.019)	.632 (.024)	.667 (.016)
Item 5	.542 (.025)	.431 (.021)	.474 (.016)	.582 (.016)
Item 6	.694 (.018)	.414 (.017)	.732 (.016)	.583 (.015)
Item 7	.543 (.019)	.459 (.018)	.841 (.014)	.331 (.015)
Item 8	.297 (.018)	.394 (.017)	.703 (.014)	.362 (.016)
Item 9	.429 (.018)	.399 (.018)	.370 (.017)	.409 (.016)
Item 10	.439 (.018)	.287 (.018)	.454 (.017)	.278 (.017)
Item 11	.409 (.019)	.456 (.017)	.581 (.016)	.411 (.021)
Item 12	.527 (.019)	.413 (.017)	.470 (.016)	N/A
Item 13	.415 (.018)	.432 (.018)	.589 (.017)	N/A
Item 14	.520 (.019)	.279 (.018)	.595 (.016)	N/A
Item 15	.438 (.021)	.128 (.025)	.487 (.018)	N/A

Table 4: *Summary of scores by cohort and subtest.*

Person Ability (θ)												
	Figural Reasoning			Figural Classification			Matrices			Reasoning		
Cohort	2	3	4	2	3	4	2	3	4	2	3	4
Min	-2.774	-2.584	-2.548	-2.472	-2.083	-2.269	-2.578	-2.263	-2.195	-1.629	-1.583	-1.517
Mean (S.E.)	0 (0)	.960 (.048)	.878 (.049)	0 (0)	1.180 (.053)	1.094 (.056)	0 (0)	.951 (.045)	.923 (.046)	0 (0)	.979 (.057)	1.119 (.060)
Max	1.608	2.160	2.198	1.856	2.707	2.793	1.628	2.114	2.098	1.994	2.811	2.860