Psychological Test and Assessment Modeling, 2023.65:414-436

DOI: https://doi.org/10.2440/001-0007

Conceptualization of the “Family Relations Reasoning Test” (FRRT)

Herbert Poinstingl[a] *& Jörn R. Sparfeldt*[b]

a Paderborn University, Faculty of Arts and Humanities
b Saarland University, Department of Educational Science

Abstract: To assess the important intelligence factor verbal reasoning, a new task type was developed. After reading a short description introducing the relations of different family members, test takers needed to identify the relation between two target family members. Based on item generation rules, 60 items of the corresponding Family Relations Reasoning Test (FRRT) were constructed. In study 1, n = 225 Austrian university students worked on one of four sets consisting of 18 German items linked with bridge items. Item Response Theory- (IRT-) analyses showed that the Rasch model held for most items, Conditional Likelihood Ratio Tests revealed strong model validity. No Differential Item Functioning (DIF) was detected, and misfit was shown for only one item. Reliability estimates and correlations with syllogism items (up to medium effect size) were partially convincing. In study 2, 60 English items were administered to n = 113 Californian college students. IRT analyses showed mostly strong validity (only five items depicted DIF and six items revealed misfit). Reliability indicators were at least acceptable, and correlations with syllogism items were of small to medium effect size. Finally, we discuss these mostly promising results suggesting the usefulness to construct rule-based items related to family relations and to assess verbal reasoning with these items.

Keywords: Family Relations Reasoning Test, verbal reasoning, crystallized intelligence, IRT analyses

Author Note

Mag. Herbert Poinstingl, Paderborn University, Faculty of Arts and Humanities
E-Mail: herbert.poinstingl@uni-paderborn.de

Prof. Dr. Jörn R. Sparfeldt, Saarland University, Department of Educational Science
E-Mail: j.sparfeldt@mx.uni-saarland.de

Introduction

Intelligence test scores are significantly correlated with and significantly predict important real-life outcomes like educational achievement and professional success (e.g., Deary, Strand, Smith, & Fernandes, 2007; Jensen, 1998; Roth et al., 2015; Schmidt & Hunter, 2004). However, these significant coefficients are far from reaching perfect relations and perfect predictions. Nevertheless and in accordance with these significant correlation and prediction coefficients, intelligence tests are widely used in scientific and applied contexts. Although a considerable number of intelligence test task types is available and well-validated (see e.g., Reynolds, Altmann, & Allen, 2021), there is a need for new and hopefully psychometrically well-functioning intelligence test task types. Regarding reasoning that is considered as central intelligence factor (e.g., Jensen, 1998) a new task type with verbal items based on item generation rules is introduced in this paper, i.e. the Family Relations Reasoning Test (FRRT). Furthermore, first psychometric results from two samples of college students dealing with evidence concerning, among others, reliability and validity aspects are reported.

Intelligence and intelligence test scores

Relying on broad consensus in the scientific community of intelligence researchers, Gottfredson (1997) defined intelligence as “a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience” (p. 13; see also Lubinski, 2000). The “ability to reason” of this definition refers to one aspect of particular importance. Since Thurstone (1938) included reasoning as an important factor of intelligence in his seven Primary Mental Abilities many intelligence theories comprise reasoning as an important part of intelligence (e.g., Carroll, 1993; Schneider & McGrew, 2012). Furthermore, reasoning is theoretically and empirically closely related to general intelligence in the sense of the *g*-factor of general intelligence or Spearman’s *g* (e.g., Jensen, 1998).

Reasoning tests are typically either stand-alone intelligence tests or important subtests of more comprehensive intelligence test batteries, including task types like number series (e.g., Liepmann, Beauducel, Brocke, & Nettelnstroth, 2012; Berndl, Steinfeld, & Poinstingl, 2012; Kersting, Althoff, & Jäger, 2008; Weiß, 2007) assessing numerical reasoning, figural analogies (e.g., Jäger, Süß, & Beauducel, 1997) and figural matrices tests (e.g., Formann & Piswanger, 1979; Becker, & Spinath, 2014; Raven & Court, 1938; Weiß, 2007) assessing figural reasoning, and verbal analogies (e.g., Jäger, Süß, & Beauducel, 1997; Kersting, Althoff, & Jäger, 2008) assessing verbal reasoning. Further reasoning test tasks require to solve mathematical problems (e.g., Kubinger & Gamsjäger, 2023).

Considering that many existing reasoning tests are well-known not only in the scientific community but also in the broad public, and that the test scores of many rule-based reasoning items can be increased substantially by repeatedly working on and practicing the items as well as by coaching and training programs (e.g., Kulik, Bangert-Drowns, & Kulik, 1984; Kulik, Kulik, & Bangert, 1984; Scharfen et al., 2018), selection and admission decisions based on correspondingly increased test scores, that are not going in hand with enhanced dispositional intelligence, might be impaired (see e.g., Schneider et al., 2020; Schneider & Sparfeldt, 2021). New task types and concepts to assess reasoning are one approach to avoid such impairments. Therefore, the construction of a new task type assessing reasoning seems to be useful and promising.

To categorize reasoning tests, Kubinger (2023) suggested the combination of one intelligence facet dimension and one content dimension. Regarding the intelligence facet dimension, Kubinger relied on the distinction between "fluid" and "crystallized" intelligence tasks as prominently emphasized by Cattell (e.g., 1963). The content dimension is based on the distinction between different contents as described by, for example, Jäger (1982) in the Berlin Model of Intelligence Structure (i.e., verbal, numerical, figural). Notably, the systematic combination of the intelligence facet dimension with the content dimension resulted in six categories of different reasoning tasks (see also figure 1 in Kubinger, 2023). For the combination of the crystallized intelligence facet dimension with the verbal (i.e., lexical) content dimension of Kubinger's suggestion, the Family Relations Reasoning Test (FRRT) was introduced. Importantly, the items were supposed to be based on item construction rules that allowed creating high item numbers with on the one hand similar and on the other hand differing characteristics.

Concept of the Family Relations Reasoning Test[1]

The concept of the FRRT is based on Kubinger's idea to assess reasoning in the verbal domain with tasks that are based on family relations. Thereby an FRRT-item consists of two parts: (1) reading a short story about the relationships of several family members and (2) finding the right relationship between two specific members by logical operations (Skoda, 2005). The verbal material of the short story describing the family relationships between two or more family members makes up the lexical facet of the items; the crystallized facet is characterized by knowledge about family relations. This can be illustrated by a simple example item: "Tim's son, Tony, has a son called Hugo. What is the relationship from Hugo to Tim?" (Answer: grandson).

[1] This idea has never explicitly been published but see Skoda (2005)

A pilot approach of such a reasoning test based on family relations (Skoda, 2005) was created by applying only two principles. The first principle is described by the “complexity of family relationships” with four levels: “nuclear family”, “relations in the second degree”, “extended family members”, “patchwork family”. The second principle was made up by the “total number of relations used in the item”. It was expected that the item difficulty increased when the complexity of family relationships and the number of relations increased. Subsequently, 100 corresponding items were constructed, partitioned in four groups, and administered to 264 students in lower Austria. However, the results of Rasch-model-analyses revealed a substantial misfit of about half of the items (49 misfitting items and 4 erroneous items; Poinstingl, 2009). Nevertheless, the approach of constructing rule-based reasoning items dealing with family relations as well as the experiences of the item construction process seemed to be promising and formed the basis for an improved version. The main improvements were related to much more specific item construction rules.

The first version of the FRRT (Schechtner, 2009; Hansmann, 2010) was characterized by a more standardized item construction process based on item generating rules. In a first step, simple item generating rules were created by considering (i) a set of objects (i.e., roles in families like daughter, sister, or wife), (ii) a set of family relations, and the systematic combination of these family relations and objects (see table 1). These item generating rules were applied (dressed up) in order to build up an item pool. The created items consisted of a short passage introducing family member (i.e., objects) and their family relations based on the item generation rules as well as a question about the family relation of two specific family members.

Specifically, the set of objects comprised different family members: father, mother, son, daughter, wife, husband, brother, sister, parents, children, grandfather, grandmother, grandson, and granddaughter. Additionally and in order to raise item difficulties, further objects (family members) were included: uncle, aunt, nephew, niece, and cousin. Further family members like great-uncle and great-aunt were not included in the item generating rules, but might be included in future extensions. For our purpose, uncle was defined as the husband of the aunt or the brother of mother or father. Aunt was defined as the wife of the uncle or the sister of mother or father. Cousin (m) was determined as the sun of aunt/uncle or the nephew of mother or father. Cousin (f) was understood as the daughter of uncle/aunt and the niece of mother or father.

In the following enumerations depicted in table 1 object X functioned as a placeholder for names. To vary the relations, the following procedure was exerted: “A is the brother of B” was changed to “B’s brother A”. For reasons of lucidity these relations were not noted in the enumeration beyond (see Table 1 for the basic relations of the item construction).

Table 1
Objects (i.e., family members) and basic relations forming the basis of the item generating rules for the Family Relations Reasoning Test (FRRT)

Object	Relation	Object	Relation
Basic Relations			
father/ husband/male	X is father of S/D	daughter/sister/ female	X is daughter of F/M
	X is husband of W		X is sister of B/S
	X has son S		X has father F
	X has daughter D		X has mother M
mother/wife/ female	X is mother of S/D	grandfather/ grandmother	X is grandfather of GS/GD
	X is wife of H		X is grandmother of GS/GD
	X has son S		X has grandson GS
	X has daughter D		X has granddaughter GD
son/brother/male	X is son of F/M	grandson/ granddaughter	X is grandson of GF/GM
	X is brother of B/S		X is granddaughter of GF/GM
	X has father F		X has grandfather GF
	X has mother M		X has grandmother GM
Enhanced Relations			
uncle	X is the uncle of Ni/Ne	aunt	X is the aunt of Ni/Ne
	X has niece Ni		X has niece Ni
	X has nephew Ne		X has nephew Ne
nephew/male	X is nephew of U/A	niece/female	X is niece of U/A
	X has aunt A		X has aunt A
	X has uncle U		X has uncle U
cousin/male	X is cousin of C	cousin/female	X is cousin of C
	X has cousin C		X has cousin C

Notes: “X” means placeholders to specified objects (i.e., names of family members).

These aspects can be illustrated easily by an item example: “Peter is the son of Cathy. Angie is the sister of Peter. Angie is the____ of Cathy.” (solution: daughter). This item was created by combining the relations with objects and including real names for the objects. The relations considered in this example item are “is son of“, “is sister of", and "is the daughter of" (required relations). The objects are A (initial object; Peter), B (mother; Cathy), and C (sister; Angie). The solution for the example item is daughter, as mentioned.

Different characteristics of the item construction were tentatively assumed to influence the item difficulty of the newly constructed items. The set of family relations summarized in table 1 comprised the "complexity of the relations" that was used while constructing items with different item difficulties. Additionally, the difficulties of the items were constituted by the considered number of relations and objects, the number of applied generations, and an increasing size of the story text. Regarding the relation that was asked for in the item, the position of the corresponding objects and the distance between the two objects in the story text might in addition influence the item difficulty. Moreover, higher item difficulties were expected for items with higher more distal relations (e.g., niece instead of daughter), the numbers of family relations needed to solve the item, the complexity of the vocabulary in the story text, or the complexity of grammar in the story text. Furthermore, including redundant information in the items might influence the corresponding item difficulty.

The items were generated based on item generating rules (IGR) that are the result of the combination of relations and objects. In the aforementioned item example, the three mentioned relations were combined with the three objects (i.e., family members) and "dressed up" by filling the place holders with corresponding names. While solving the item, the two relations mentioned in the item need to be considered and combined in order to find the third relation to solve the item correctly (further relations not needed to solve the item were not included in the story text of this item, but could easily be added).

Relying on these item generation rules and the procedures mentioned above, 64 FRRT items were created. Importantly, the unambiguous identification of the gender of the family members (i.e., objects) is crucial to solve the items successfully. To conclude, this rule-based procedure constitutes the basis of the item universe and, thereby, the creation of structurally parallel items as well as items with varying item difficulties. Finally, the item difficulty levels were expected to vary systematically with certain item characteristics, as mentioned before.

The main aim of the present study was to investigate important characteristics of the new Family Relations Reasoning Test in college/university student samples. In two separate studies, we described a first evaluation of the usability of the item construction rules. In study 1, FRRT items were administered in German to a sample of Austrian university students; in study 2, FRRT items were administered in English to an US-American student sample. Relying on an IRT framework, we inspected the accordance of the FRRT items with the one-dimensional Rasch model (Rasch, 1960) and whether the observed data were better specified by subgroup-specific than general parameter estimates (Conditional Likelihood Ratio Test – CLRT, Andersen, 1973; DIF-analyses – Wald type test, Glas & Verhelst, 1995). Furthermore, first reliability evidence was inspected by relying on person separation reliability (r_{sep}) and Cronbach's α. Furthermore, we analysed the relations between the FRRT items and some syllogism items to get first evidence related to convergent validity.

Study 1

The main aim of study 1 was a first evaluation of the Family Relations Reasoning Test (FRRT). The items were constructed based on the mentioned item construction rules. In this study, we analysed mainly whether the corresponding IRT model held for these items. Furthermore, first evidence regarding reliability and convergent validity was inspected.

Study 1: Methods

Procedure

The investigation took place in the laboratories of an Austrian university in Vienna in 2009 using a test administration server system (with software specifically written for this purpose; Poinstingl, 2009). Advanced psychology students were specifically trained and administered the assessment procedure in a standardized manner. There was no time limit at all. The complete assessment procedure lasted about 45 minutes (including all tests and demographics).

Sample

The convenience sample consisted of n = 225 psychology students (n = 170 female, n = 55 male; age: M = 24.88 years, SD = 5.04 years). The mother tongue of the majority of these students (92.4%) was German; all students had at least good command of the German language as required for participation in the study program. The majority of these students studied psychology and received partial credit required within the study program (i.e., participating in empirical research projects). Nevertheless, participation was voluntary.

Instruments

Verbal reasoning was assessed with $k = 64$ FRRT items constructed in accordance with the item construction rules mentioned above.[2] However, 4 out of these 64 items (51, 55, 57, 59) had to be excluded from the item pool because of serious faults. Because of practical constraints limiting the number of FRRT-items to be administered in one test session, four item sets consisting of 18 FRRT-items were created (with systematically varying item difficulties) and a linked item test design was implemented. Linkage was accomplished by bridge items (10, 14, 18, 25, 30, 42, 48; see Table 2). The students were randomly assigned to one of four groups ($n = 56/57/56/56$) working on one of the four item sets lasting on average 30 minutes (no time limit). These items were administered in two different response formats (multiple choice; five sequentially presented statements to be judged as either "true" or "false" with terminated presentation of the statements following the first "true" judgement; see Figures 3 and 4). Every item was presented in both formats, although to different students; the students of one group worked on the respective items with mixed response formats. Regarding the scoring of the multiple choice items, credit was given (coded as 1), if only the correct choice was marked, otherwise no credit was given (coded as 0). Regarding the sequential response format items, credit was given for each item (coded as 1), if the first statement marked as "true" appeared for the correct statement, otherwise no credit was given (coded as 0).

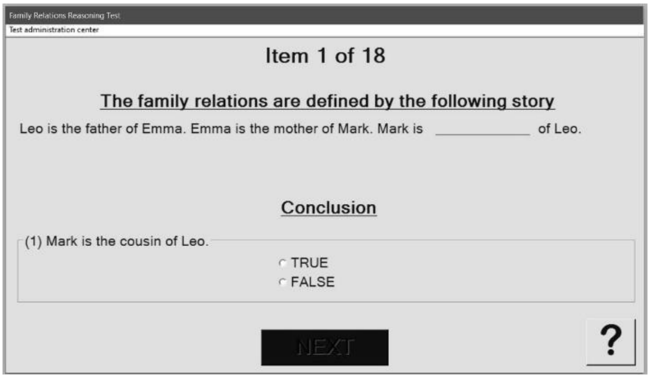
Additionally, verbal reasoning was assessed with ten syllogism items (Srp, 1994), lasting on average eight minutes (no time limit). Socio-demographic characteristics (gender, age, mother tongue) were collected, besides further variables not relevant for the present study.

Table 2
Test design with bridge items for 4 groups and 18 items (bridge items are printed in bold)

Item	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17	f18
Group 1	8	6	17	**14**	40	25	**48**	3	**30**	41	56	**22**	27	60	31	37	44	62
Group 2	2	4	**18**	**14**	**10**	24	**48**	7	**30**	45	58	23	28	54	19	34	49	61
Group 3	1	**10**	20	**14**	16	25	46	11	39	**42**	53	9	33	**48**	13	32	38	64
Group 4	5	26	**18**	**14**	29	35	43	12	21	**42**	52	15	47	**48**	**22**	36	50	63

[2] For research purposes, the items are available upon request from the first author.

Figure 3
Administration of the FRRT with sequential response format



Item 1 of 18

The family relations are defined by the following story

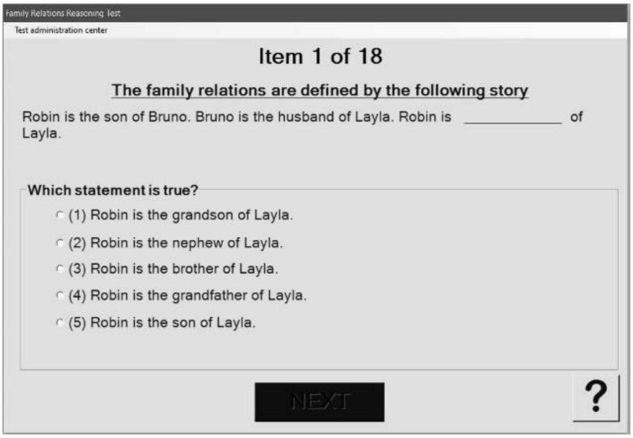
Leo is the father of Emma. Emma is the mother of Mark. Mark is ____________ of Leo.

Conclusion

(1) Mark is the cousin of Leo.

○ TRUE
○ FALSE

Figure 4
Administration of the FRRT with MC-format (“1 out of 5”)



Item 1 of 18

The family relations are defined by the following story

Robin is the son of Bruno. Bruno is the husband of Layla. Robin is ____________ of Layla.

Which statement is true?

○ (1) Robin is the grandson of Layla.
○ (2) Robin is the nephew of Layla.
○ (3) Robin is the brother of Layla.
○ (4) Robin is the grandfather of Layla.
○ (5) Robin is the son of Layla.

Statistical analyses

Statistical analyses and model analyses were done with the statistical package R V3.6.1 (R Core Team, 2007) and the following R-Packages: eRm: Extended Rasch Modeling. V1.0-0 (Mair & Hatzinger, 2007), psych: Procedures for Personality and Psychological Research. V1.8.12 (Revelle, 2007).

Regarding the model fit of the FRRT items and test, IRT analyses were performed by the application of the Rasch model (1-PL model; Rasch, 1960). The Likelihood Ratio Test (LRT) provided an examination of the total model. Conditional Likelihood Ratio

Tests (CLRT; Andersen, 1973) were applied to check whether the observed data could be better specified by parameter estimates taken from specific subgroups rather than the parameter estimates taken from the total data set (e.g., Kubinger, 1989). Subgroups were based on median-splits (FRRT score, age, gender, and the response format of the FRRT items). Concerning the DIF analyses of the items, Wald type tests (Glas & Verhelst, 1995) were performed. We tested whether the fit of a general Rasch model with general item parameters for the whole sample or subgroups (with different item parameters for the subgroups) provided a better fit.

Finally, infit statistics (Bond & Fox, 1997; Mair & Hatzinger, 2007) were analysed to check the itemfit of the FRRT. Infit statistics are standardized values in order to detect items with substantial deviations from the model. Items with values outside $-2 < Z < 2$ are considered to significantly ($\alpha = .05$) deviate from the Rasch model; thus, we removed these items from the item pool. The mentioned analyses were run with the software package eRm (Mair & Hatzinger, 2007).

Regarding reliability evidence of the FRRT, Cronbach's α and the person separation reliability (r_{sep}) were inspected. The concept of person separation reliability (defined as the proportion of person variance that is not due to error) is very similar to reliability indices such as Cronbach's α (Mair, 2020). α and r_{sep} should reach at least .7 (Wright & Stone, 1999).

Regarding convergent validity evidence, we analysed the correlations between the FRRT and the sum score of the Syllogism-items.

Study 1: Results

Concerning the IRT model analyses, three items (1, 9, 13) had to be excluded from the item estimation process because of full item scores (i.e., all students solved these three items correctly). The CLRTs ($\alpha_{crit} = .01$) with split criteria achievement (median), gender, age (mean), and response format showed no statistically significant result (Table 3).

Table 3
Study 1: Results of the CLRT using the split criteria FRRT achievement score (median), gender, age (mean), and response format.

Criteria	χ^2	χ^2_{crit} (α = .01)	*df*	*p*
Achievement score (median)	33.09	46.96	27	.19
Gender	43.09	72.44	47	.64
Age (mean)	62.47	79.84	53	.18
Response Format	48.69	73.68	48	.45

Notes: The lower number of degrees of freedom for the criterium "FRRT score (median)" was caused by the fact that some very easy items were solved correctly by all high achievers and that some very difficult items were not solved correctly by a least one lower achiever.

Regarding possible Differential Item Functioning (DIF) for FRRT achievement, gender, age, and response format, the results of the Wald type tests did not reach significant resuls (based on $p < .01$). Only one item (46) exceeded the lower boundary indicating item misfit ($Z = -2.03$); the infit of the remaining items reached values within the mentioned boundaries ($-2 \leq Z \leq 2$). Further information about the validity (graphical model checks) can be found in the appendix (Figure 5, Figure 6, Figure 7). Valid items are located near the 45-degree-line. In the R-Package eRm (Mair & Hatzinger, 2007) the continuous control lines below and above most items indicated non-significant deviations from the model.

Concerning reliability evidence, person separation reliability was, however, rather low ($r_{\text{sep}} = .61$). For the four item sets respectively the four groups (group 1/2/3/4), Cronbach's α reached α = .63/.65/.64/.69 (mean: $\alpha = .65$, $SD = .03$). Regarding the relations of the FRRT with the Syllogism-test as indicator of convergent validity, the following correlations were found for the four groups: r = .41/.01/.12/.39.

Study 1: Summary

In summary, the IRT-based model checks assured strong model validity for the German FRRT (after excluding 3 out of 60 items) These results were confirmed by DIF analyses revealing no hints for differential item functioning. The evaluation of item fit showed only one item with misfit. The reliability indicators revealed, however, rather low reliability coefficients of the FRRT (adhering to DeVellis, 2017). These rather low reliability estimates call for further research in order to improve the reliability of the FRRT to reach evidence indicating sufficient reliability. Regarding convergent validity, the correlations of the FRRT and Syllogism-items were of moderate effect size in two groups and negligible for two other groups (adhering to Cohen, 1988). This inconsistency calls for further research. Taken together, the results based on IRT analyses are very convincing and promising, whereas the (further) evidence regarding reliability and validity are calling for further research. One particularly interesting aspect possibly improving the application of the instrument in English-speaking samples is the construction of an English FRRT version.

Study 2

The main aim of study 2 was to introduce and empirically inspect the English version of the FRRT. Therefore, we genereted 60 items for the English version of the FRRT by applying the same item generating rules used for the construction of the German version. Particular attention was paid to ensure equivalence between the English and the German version of the FRRT items in terms of vocabulary, syntax, semantics, and the names of the family members (i.e., objects). In study 2, the main focus was on analysing whether the unidimensional Rasch-model holds for these items. Additionally, first evidence regarding reliability and convergent validity was aimed for.

Study 2: Methods

Procedure

The investigation took place in fall 2009 in the laboratories of a college in California (USA) using the same test administration procedure as in study 1.

Sample

The sample comprised *n* = 113 students (*n* = 70 female, *n* = 43 male; age: *M* = 24.37 years, *SD* = 5.99 years). The mother tongue of the majority of these students was either English (*n* = 82, 72.6%) or Spanish (*n* = 23, 20.4%), besides single individuals with another mother tongue. All participants had excellent English language skills. Participation was voluntary. The students were not graded or rewarded for their participation.

Instruments

Verbal reasoning was assessed with English versions of the k = 60 FRRT items from study 1, constructed in accordance with the item construction rules mentioned above. Special attention was paid to construct equivalent English items of the FRRT in terms of vocabulary, syntax, semantics, and the names of the family members. However, only the multiple-choice response format was used in study 2. As in study 1, the same four item sets with the same linked items design were constructed; and the students were randomly assigned to work on one of the four item sets (n = 28/31/27/27; see table 2). Working on these items lasted on average 30 minutes (no time limit).

As in study 1, verbal reasoning was assessed with ten syllogism items (Srp, 1993), lasting on average eight minutes (no time limit). Furthermore, socio-demographic characteristics (gender, age, mother tongue) were assessed, besides further variables not relevant for the present study.

Statistical analyses

The same statistical analyses as in study 1 were performed in study 2. Rasch model analyses of the English version of the FRRT were tested with LRT. CLRT was run with the split criteria FRRT achievement (median-split of the test score) and age (median-split), but was not applicable for gender and response format.

Study 2: Results

The CLRT model checks showed no statistically significant result for the items of the English FRRT version for both split criteria (Table 4).

Table 4
Results of the Conditional Likelihood Ratio Test using the split criteria achievement (median of test score) and age (mean) in study 2

Criteria	χ^2	$\chi^2_{crit\ (\alpha = .01)}$	*Df*	*p*
Achievement score (median)	40.85	58.62	36	.27
Age	60.60	85.95	58	.38

Regarding the detection of posssible Differential Item Functioning (DIF) for both used criteria (achievement, age), by running corresponding Wald tests, an itemwise inspection revealed for most items no significant evidence for DIF; however, evidence for DIF was shown for five items (regarding FRRT achievement: Item 4 – z = 3.29, $p < .001$, Item 20 – z = 3.10, $p < .001$, Item 38 – z = 2.75, $p = .006$, Item 46 – z = 3.55, $p < .001$; regarding age: Item 64 – z = 2.75, $p = .005$). Infit statistics indicated for most items no significant evidence regarding infit. However, the results revealed lacking item fit for six items: Items 4 ($z = 2.67$), 25 ($z = 2.87$), and 42 ($z = 3.31$) exceeded the upper boundary of 2.0, whereas Items 13 ($z = -2.37$), 24 ($z = -2.09$), and 27 (z = -2.56) fell below the lower boundary of -2.0.

Concerning reliability evidence, person separation reliability reached $r_{\text{sep}} = .86$ (based on all 60 items because of the non-significant CLRT result). For the four item sets/groups (1/2/3/4), Cronbach’s alpha values of α = .86/.86/.82/.92 were shown ($\alpha_{mean} = .86$, $SD = .04$). Regarding the relations between the FRRT and the syllogism scores indicating convergent validity, correlations reached r = .29/.35/.12/.11.

Study 2: Summary

In summary, the IRT-based model checks assured mostly strong model validity for the new 60 English FRRT items. These results were mainly confirmed by DIF analyses (with no hints for differential item functioning for 55 items) and infit analyses (with no hints for infit for 54 items). Additionally, graphical model checks can be found in the appendix (Figure 8, Figure 9). Regarding the English FRRT version, the reliability indicators revealed at least acceptable values (adhering to DeVellis, 2017). Concerning convergent validity in terms of the relations between the FRRT and Syllogism-items, the correlations were of (almost) medium effect size (adhering to Cohen, 1988) in two groups and negligible in two other groups. This inconsistency calls for further research, as well. Taken together, these results for the English FRRT version are mostly promising, as well. Notably, we have to keep in mind that the sample size was rather low. Correspondingly, further research and replications in studies with larger samples are needed.

General Discussion

The main aim of both studies was an evaluation of the Family Relations Reasoning Test (FRRT) as a new task type and instrument to assess verbal reasoning; the analyses focussed on IRT-analyses as well as evidence regarding reliability (i.e., person separation reliability, Cronbach's alpha) and convergent validity (i.e., correlations with syllogism items). Based on item construction rules, corresponding German FRRT items were constructed and inspected in Austrian university students (study 1); additionally, corresponding English FRRT items were investigated in an US American college student sample.

Concerning study 1 with Austrian university students, the IRT analyses revealed that the Rasch model held for most items. Only 3 out of 60 items had to be excluded from the estimation processes because of full item scores. CLRT model checks depicted strong model validity for the German FRRT version applying four criteria (FRRT achievement, gender, age, and response format). These CLRT results were confirmed by DIF analyses that showed no differential item functioning. The evaluation of item fit revealed only one item (46) with misfit. To conclude, these IRT results concerning our main criterion (Rasch model analyses) were convincing. However, the results related to reliability evidence showed the need for further research. The correlation coefficients between the FRRT and the syllogism items as indicators for convergent validity were satisfactory in groups 1 and 4 (r = .41/.39), but not in groups 2 and 3 (r = .01/.12). The students were randomly assigned to the four groups and item sets; however, this result pattern calls for further research.

Concerning study 2 with Californian college students working on English FRRT items, the Rasch model analyses showed that the Rasch model held for these items. Nevertheless, DIF analyses showed DIF for 5 out of 60 items and infit analyses indicated misfit for 6 out of 60 items. To conclude, these IRT-based analyses were promising for most English FRRT items. In partial contrast to the results in study 1 with German items and samples, the evidence regarding reliability ($r_{sep} = .86$; $.82 \leq \alpha \leq .92$) was satisfactory in all four groups/item sets. Regarding the correlations with the syllogism items as indicators for convergent validity the coefficients reached small to medium effect size in groups 1 and 2 (r = .29/.35); however, the (very) small effect size of these correlations in groups 3 and 4 (r = .12/.11) calls for further research.

Regarding the Rasch model analyses the promising results in both studies indicate that the item construction rules and the realized rule-based item construction allow developing verbal reasoning items based on family relations. Furthermore, this rule-based procedure allows the construction of further FRRT items with differing item difficulties, either to construct structurally parallel further items with similar item difficulty values, much easier items, or even much more difficult items. Thereby, the construction of further items with adequate fit for more or less able samples under investigation (without, for example, ceiling or bottom effects) should be easily possible. Furthermore, the promising results that the Rasch model held for (most of) the items are promising regarding adaptive testing tools in order to come to, for example, very timeefficient instruments. In future studies, it seems fruitful to investigate the correspondence between a priori assumed item difficulties and empirical item difficulties.

Related to the reported reliability coefficients as well as the convergent validity coefficients, the results are, however, more equivocal. Whereas the reliability estimates of the English FRRT version are convincing (r_{sep} = .86; .82 ≤ α ≤ .92) the corresponding results of the German FRRT version indicate a need for further improvements (r_{sep} = .61; .63 ≤ α ≤ .69). Because the number of items of the four test sets was rather low, increasing the number of the items might go in hand with higher reliability coefficients. To estimate the (predicted) reliability of correspondingly lengthier tests, we used the well-known Spearman-Brown formula resulting in more satisfactory results for test sets with 27 items instead of 18 items, that is increasing the item numbers substantially to one and a half times the original number (α = .71/.74/.73/.77). Furthermore, one could eliminate a few items that contribute to the low reliability estimates and supplement the instrument by new items to hopefully reach higher reliability estimates. These estimated values are (more) consistent with the results of the IRT analyses discussed above. Similarly, whereas the correlation coefficients of the FRRT with the syllogism items indicating convergent validity were almost satisfactory in two of four groups (for the German as well as the English version), the irritating low coefficients for the remaining two groups call for further research. Additionally, further validity evidence in terms of, for example, relations with other variables and instruments are needed to strengthen the nomological network of the FRRT in terms of convergent (e.g., further intelligence indicators, achievements in school and university) and discriminant validity (e.g., personality factors like conscientiousness and extraversion). Finally, further research is needed regarding different (larger and more divers) samples as well as samples from different cultures and nations in order to strengthen the first results of our study indicating the usability of the English as well the German FRRT version.

To conclude, we described the new verbal reasoning test task concept, the item construction process, and first results regarding the test evaluation of the German and the English versions of the Family Relation Reasoning Test. In future studies, a thorough examination of items with the Linear Logistic Test Model (Fischer, 1973) is highly recommended in order to reach an even more ambitious goal: an automatic item generation process for creating a number of valid items with varying item difficulties with predicted characteristics (like item difficulty) for adaptive testing procedures.

References

Andersen, E. B. (1973). A Goodness of Fit Test for the Rasch model. *Psychometrika, 38,*123–140.

Becker, N., & Spinath, F. (2014). *Desigma – Advanced – Design a Matrix – Advanced*. Göttingen: Hogrefe.

Berndl, G., Steinfeld, J. & Poinstingl, H. (2012). *Schlussfolgerndes Denken numerisch: Der Wiener Zahlenreihentest* [Numeric Reasoning. The Viennese number series test]. In K. D. Kubinger, M. Frebort, L. Khorramdel, & L. Weitensfelder (Eds.), *Self-Assessment: Theorie und Konzepte* (Spp. 161-169) [Self-Assessment: Theory and Concepts (Spp. 161-169)]. Lengerich: Pabst.

Bond, T., & Fox, C. (1997). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, NY: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). New York: Erlbaum.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21.

DeVellis, R. F. (2017). *Scale Development. Theory and Application* (4th ed). Los Angeles: Sage.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT)* [Viennese Matrices-test, WMT]. Weinheim: Beltz Test.

Glas C.A.W, Verhelst N. (1995). Tests of Fit for Polytomous Rasch Models. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications*, pp. 325–352. New York: Springer.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13–23.

Hansmann, B. C. (2010). *About the psychometric quality of various multiple choice response formats in the context of cultural differences between Austria and the United States of America* [unpublished master thesis]. Universität Wien.

Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test (BIS), Form 4 [Berlin Intelligence Structure-Test (BIS), Form 4]*. Göttingen: Hogrefe.

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

Kersting, M., Althoff, K. & Jäger, A. O. (2008). *WIT-2. Der Wilde-Intelligenztest 2*[Wilde Intelligence Test 2]. Göttingen: Hogrefe

Kubinger, K. D. (Ed.). (1989). *Moderne Testtheorie: ein Abriß samt neuesten Beiträgen* [Modern Test Theory: A Brief Survey, With Recent Contributions]. Weinheim: Beltz.

Kubinger, K.D. & Gamsjäger, C. (2023). Conceptualization of a new Two-way Figural Reasoning Test. *Psychological Test and Assessment Modeling,* 339-353.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179–188.

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21*, 435–447.

Liepmann, D., Beauducel, A., Brocke, B., & Nettelnstroth, W. (2012). *IST-Screening. Intelligenz-Struktur-Test-Screening [IST-Screening. Intelligence-Structure-Test-Screening].* Hogrefe.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points". *Annual Review of Psychology, 51*, 405–444.

Mair P, Hatzinger R (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software,* 20(9). http://www.jstatsoft.org/v20/i09/.

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2020). Package 'eRm'. *R Foundation, Vienna, Austria*.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Poinstingl, H. (2009a). *Der Remote Testing Approach*. [Paper presentation]. General Online Research 2009. Vienna.

Poinstingl, H. (2009b). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychological Test and Assessment Modeling, 51*, 123.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research: Copenhagen.

Raven, J. C., & Court, J. H. (1938). *Raven's progressive matrices*. Los Angeles: Western Psychological Services

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reynolds, C. R., & Livingston, R. A. (2021). *Mastering modern psychological testing*. Springer International Publishing.

Revelle, W. (2022). *psych: Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://cran.r-project.org/package Version = 2.2.9.

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003.

Schechtner, C. M. A. (2009). *Entwicklung eines rationalen Itemkonstruktionsprinzips als Basis eines sprachlichen Reasoning-Tests [Development of rational item construction principle as the base of a lexical reasoning test]*. [unpublished master thesis]. Universität Wien.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162–173.

Schneider, B., Becker, N., Krieger, F., Spinath, F. M., & Sparfeldt, J. R. (2020). Teaching the underlying rules of figural matrices in a short video increases test scores. *Intelligence, 82*, 101473.

Schneider, B. & Sparfeldt, J.R. (2021). How to solve number series items: Can watching video tutorials increase test scores? *Intelligence, 87*, 101547.

Schneider, W.J., & McGrew, K.S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). New York: Guilford Press.

Skoda, S. (2005). Der Verwandtschaften-Reasoning-Test. Konzept und erste empirische Erprobung [The Relationship Reasoning Test. Concept and first empirical evaluation]. [unpublished master thesis]. Universität Wien.

Srp, G. (1994). *Syllogismen* [Syllogisms]. Frankfurt/M.: Swets Test Services.

Süß, H.-M. (2003). Intelligenztheorien [Intelligence theories]. In K.D. Kubinger & R.S. Jäger (eds.), *Schlüsselbegriffe der Psychologischen Diagnostik [Key terms of psychological assessment]*. (Spp.217–224). Weinheim: Beltz.

Thurstone, L. L., & Thurstone, T. G. (1938). *Primary mental abilities* (Vol. 119). Chicago: University of Chicago Press.

Weiß, R. H. (2007). *Grundintelligenztest Skala 2-Revision (CFT 20-R): mit Wortschatztest und Zahlenfolgentest-Revision (WS/ZF-R)* [Test for Basic Intelligence Scale 2 - Revision (CFT 20-R): with vocabulary test and number series test - Revision (WS/ZF-R)]. Göttingen: Hogrefe.

Wright, B.D., & Stone, M.H. (1999). *Measurement essentials.* Wilmington: Wide Range.

Acknowledgement

We are thankful to Mag. Darina Lavtarova, MA for the very careful review of all FRRT items.

Appendix

Table 5

Family Relations Reasoning Test (English version, 5 examplary items and item generation rules)

nr	dressed up items	item generating rule	correct relation
2	Robin is the son of Bruno. Bruno is the husband of Layla. Robin is ____________ of Layla.	A is the son of B. B is the husband of C. A is ____________ of C.	Robin is the son of Layla
10	Phil is the brother of Aidan. Aidan is the father of Billy. Billy is ____________ of Phil.	A is the brother of B. B is the father of C. C is ____________ of A.	Billy is the nephew of Phil
30	Edwin has an uncle called Joe and a mother called Doris. Amber is the wife of Joe and the mother of Mitch. Linda is the sister of Edwin. Amber is ____________ of Linda.	A has an uncle called B and a mother called C. D is the wife of B and the mother of E. F is the sister of A. D is ____________ of F.	Amber is the aunt of Linda
43	Tara is the mother of Paige. Hazel is the mother of Alan. Dan has a wife called Hazel. Elias has a sister called Paige. Alan is the husband of Tara. Elias is ____________ of Dan.	A is the mother of B. C is the mother of D. E has a wife called C. F has a sister called B. D is the husband of A. F is ____________ of E.	Elias is the grandson of Dan
58	Owen is the father of Bella. Nina has a son called Tim and a husband called Felix. Mia is the sister of Lance. Bella has a brother called Felix. Lance is the father of Ralph and has a wife called Bella. Tim is ____________ of Ralph.	A is the father of B. C has a son called D and a husband called E. F is the sister of G. B has a brother called E. G is the father of H and has a wife called B. D is ____________ of H.	Tim is the cousin of Ralph
61	Eric's mother Karen is the niece of Noah's brother Noel. Cora is the daughter of Tina. Sandy and Cora are sisters. Tina has a husband called Noel. Ben is the brother of Sandy. Ben is ____________ of Karen.	A's mother B is the niece of C's brother D. E is the daughter of F. G and E are sisters. F has a husband called D. H is the brother of G. H is ____________ of B.	Ben is the cousin of Karen

Figure 5

Study 1 – Graphical Model Check (Achievement score)

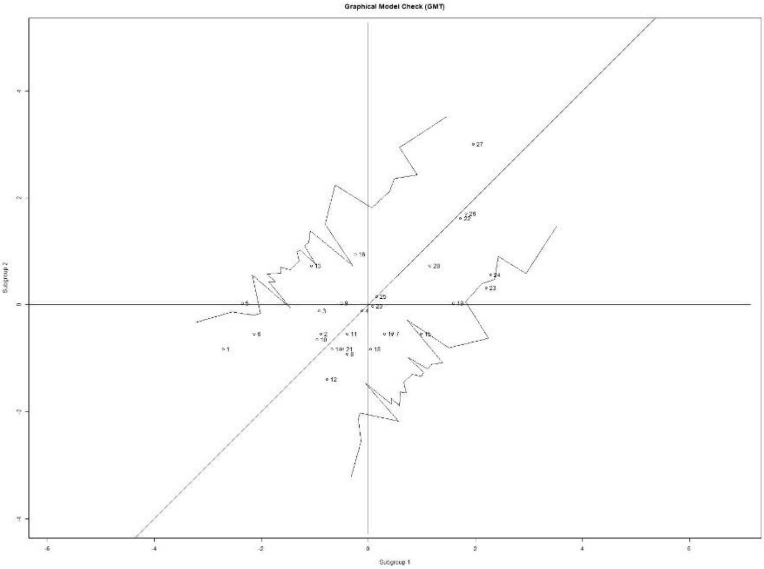


Figure 6

Study 1 – Graphical Model Check (Criterion Gender)

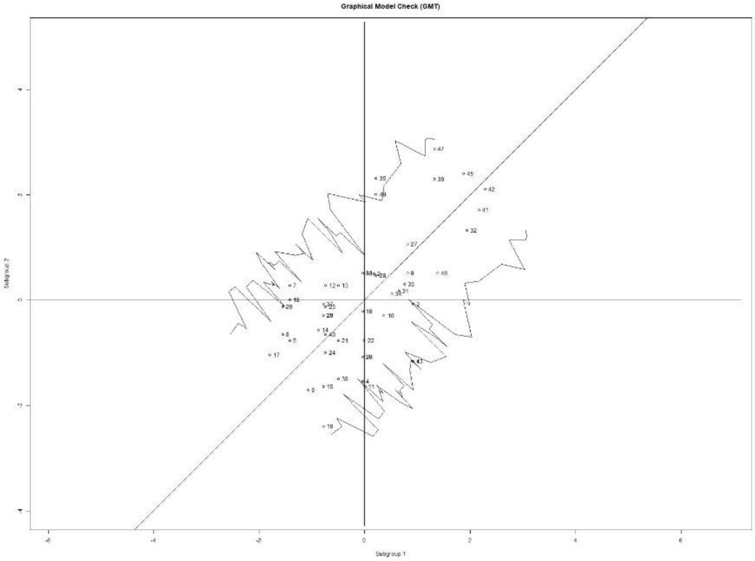


Figure 7

Study 1 – Graphical Model Check (Criterion Age)

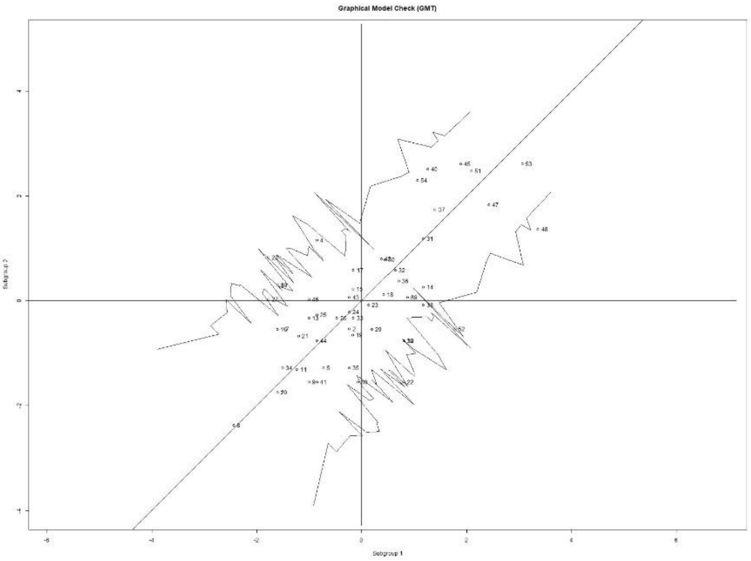


Figure 8

Study 2 – Graphical Model Check (Achievement score)

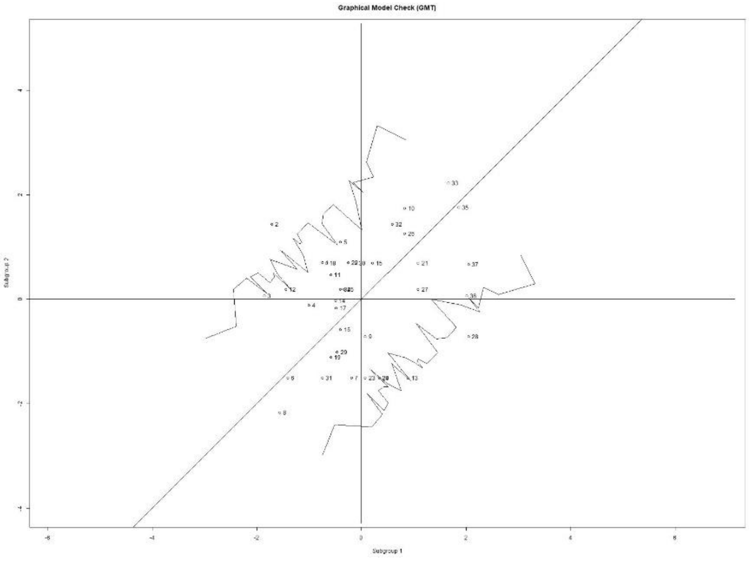
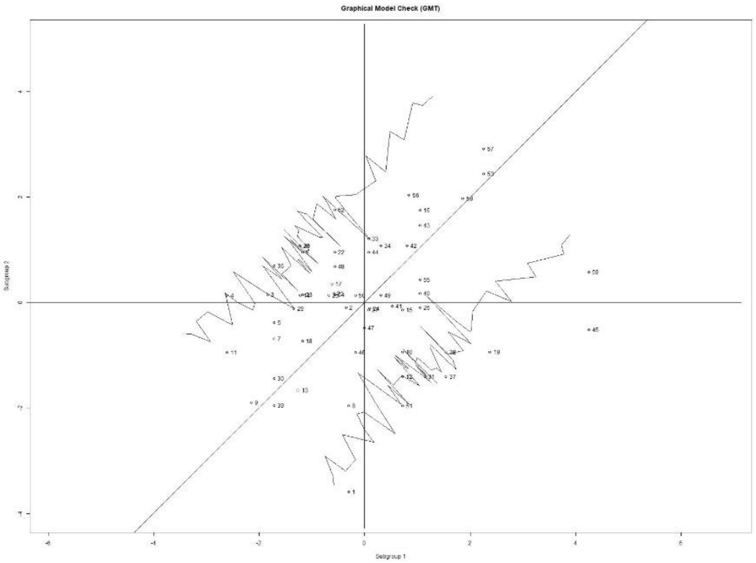


Figure 9
Study 2 – Graphical Model Check (Criterion Age)


Graphical Model Check (GMT)
Subgroup 1
Subgroup 2