Psychological Test and Assessment Modeling, 2023.65:402-413
DOI: https://doi.org/10.2440/001-0006

Conceptualization of the Reasoning-Test "Reality-contradicting Syllogisms"

Sarah Treiber & Klaus D. Kubinger

University of Vienna

Abstract: Considering so-called reasoning tests, almost only test concepts with figural item contents are practically in use – this being in *Raymond B. Cattell*'s tradition of aiming for culture-fair tests. However, as Kubinger (2023a) recently suggested there are six categories of reasoning tests, i.e. a two times three classification of *fluid* vs. *crystallized* facets and *lexical* vs. *numerical* vs. *figural* contents. And especially the combination of *fluid* facets with *lexical* contents is hardly available to a practitioner. Though, there is at least the approach to use the formal-logical principle of a "syllogism", which will be re-activated in this paper. In contrast to a lot of Srp's items of her test (Srp, 1994) now each item is constructed in such manner that at least one premise as well as the conclusion contradict the actual facts (the material truth) or, alternatively, it has no reference to reality by using meaningless, freely invented "words" as acting terms. A first draft of such a test *Reality-contradicting Syllogisms* with 20 items has been psychometrically analyzed according to the Rasch model. Although only two of the items have to be deleted in order to achieve *a-posteriori* model conformity, no insight could be gained about which specific components in the composition of an item are problematic so that it captures something different from the other items. For now, these items' non-conformity with the Rasch model is simply be attributed to chance. But for deeper understanding the challenge of syllogisms as a psychological test's task, further research is needed. Some hints for doing so are given in this paper.

Keywords: Reasoning, syllogism, fluid vs. crystallized facets, multiple-choice response format, Rasch model

Author Note

Prof. Klaus D. Kubinger, PhD, MSc. c/o University of Vienna, Faculty of Psychology. klaus.kubinger@univie.ac.at

Introduction

As the principle of a "syllogism" due to *Aristotle*'s approach of logic proves by definition whether a person's formal-logical power of deductive reasoning is high or low, its use as the task of a psychological (reasoning) test is obvious – a syllogism arises when two true premises unequivocally imply a certain conclusion. Hence, Srp (1994) published such a test for giving it on practitioners' disposal, which however was soon taken from the market because of the publisher's shutting down. Nevertheless, that test *Syllogisms* serves as an excellent example and reference. Produced at the very beginnings of computerized psychological tests it even applied adaptive tailored testing. And it used as the first (and up to now only) test the so-called sequential response format, meaning that the answer options are presented one after the other, as long as the testee decides the current given option is wrong; in comparison with the commonly used simultaneous presentation of all answer options at once the probability of lucky guessing a multiple choice item is thus substantial reduced (cf. Kubinger, 2019). As required for adaptive testing (cf. Kubinger, 2016), the test items were found to conform to one of the IRT (*Item Response Theory*) models, namely the Rasch model: *a-posteriori* validness of the model was given after only five of 80 items have been deleted.

We now introduce an attempt to renew the respective test conceptualization. Matter-of-factly a specific test version with only a few items is currently, since 2011, in practical use, as a part of some online self-assessment for applicants of *Computer Science & Electrical Engineering* studies at the University of Technology in Vienna (*Wiener Self-Assessment Informatik/Elektrotechnik*© 2011[1]; see Treiber, 2013).

Although, the renewed test conceptualization has not yet approached adaptive testing and furthermore has abandoned, for the the time being, the sequential response format on the conventional multiple-choice response format's behalf, it establishes an essential improvement: In contrast to a lot of Srp's items of her test *Syllogisms* now each item is constructed in such manner that at least one premise as well as the conclusion contradict the actual facts (the material truth) or, alternatively, it has no reference to reality by using meaningless, freely invented "words" as acting terms. This is due to preventing testees from simply accessing their factual knowledge about the relationships between the various terms and thus arriving at the solution without deductive reasoning.

Method

Characterizing a (so-called categorical) syllogism formally more precisely it consists of two premises (p1, p2), each of them contains a statement which concerns the relation between a first term occurring in both premises, i.e. the middle term (M), and

[1] http://studienwahl.tuwien.ac.at

premise-specific second terms (subject term S and predicate term P, respectively). And there is the conclusion (c), which deduces an unequivocal statement about the relation between both the second terms (S and P). For example:

Premise 1:	*All flowers are beautiful.*
Premise 2:	*All roses are flowers.*
Conclusion:	*All roses are beautiful.*

In this example "flowers" represents the middle term M, "roses" the subject term S, and "beautiful" the predicate term P. Via the middle term M the conclusion results.

There generally occur four different types of the proposition of two terms' relation: *universally affirmative* (i.e. *All A are B*), *particular affirmative* (i.e. *Some A are B*), *universal negative* (i.e. *All A are not B/No A is B*), and *particular negative* (i.e. *Some A are not B*). Using the formal logic notations, these four types of proposition can be written as follows – also the set theory's expressions are given:

proposition	formal logic	set theory
1 *All A are B*	$\forall x[A(x) \rightarrow B(x)]$; $\neg\exists x[A(x) \wedge \neg B(x)]$	$A \subseteq B$
2 *Some A are B*	$\exists x[A(x) \wedge B(x)]$	$A \cap B \neq \varnothing$
3 *All A are not B/No A is B*	$\neg\exists x[A(x) \wedge B(x)]$; $\forall x[A(x) \rightarrow \neg B(x)]$	$A \cap B = \varnothing$
4 *Some A are not B*	$\exists x[A(x) \wedge \neg B(x)]$	$A \not\subset B$

Meaning of the symbols: $\forall$, "all"; x, object(s); $A(x)$, object group A with object(s) x; $\rightarrow$, "leads to"; $\neg$, negation; $\exists$, "there is"; $\wedge$, conjunction ["as well as"]; $\subseteq$, subset [left set contained in right set]; $\cap$, intersection [common set of left and right set]; $\varnothing$, empty set; $\not\subset$, no subset [left set not contained in right set]

The reference to the formal logic notations of the given propositions is at least of importance as concerns the quantifier "some". In contrast to its use in everyday language in the sense of "several, albeit a few, but not all" (which does not usually include the case of "one"), the meaning of "some" with respect to the formal logic and hence with respect to a syllogism is: "there exists at least one of the subject term objects fulfilling the respective statement, but maybe even all these objects do". That is, in future it might be preferable to use consequently the expression "there is at least one" instead of the quantifier "some" – as is sometimes the case with the first draft of the test *Reality-contradicting Syllogisms* anyway. Additionally, we gave the instructive hint: *"Some" means "at least one", that is it might also be "all"*.

In two premises (p1, p2) there are four different sequence combinations with respect to the middle term, the subject term S, and the predicate term P (i.e. MP-SM, PM-SM, MP-MS, PM-MS – S and P being likewise interchangeable), which can be combined with the four different propositions in each of these premises as well as in the conclusion (c); that is, $4 \times 4^3 = 4^4 = 256$ "modi" of thinkable syllogisms result, though only

19 are indeed conclusive (cf. Freytag-Löringhoff, 1966). They all are listed in the Appendix.

When using syllogisms as the task of a psychological test, of course, in terms of statement content there are only four different response options, corresponding to the four different propositions. In order to do not risk a rather too high chance of lucky guessing with a multiple choice response format "1 out of 4" (a single answer option out of four is correct), the test *Reality-contradicting Syllogisms* in the here presented first draft conceptualizes a fifth answer option, which always is just a reformulation of one of the wrong answer options (e.g. *All A are not B* being changed to *No A is B*).

The now given item-pool contains more or less frequently everyday words or invented "words" for the acting terms, but, as already indicated, all items contradict the actual facts (the material truth) or have no reference to reality, respectively. Some items' premise(s) and their conclusion use the conjunctive mood, because, according to Srp (1994), this makes items more difficult. Figures 1 to 3 give three examples, the first a most simple one of the type MP-SM with the propositions for p1-p2-c: 1-1-1. The second example is of type MP-MS with the propositions for p1-p2-c: 1-1-2. And the third example is of type PM-SM with the propositions for p1-p2-c: 1-3-3; there the acting terms are freely invented, and also the conjunctive mood is used as well as once the formulation "*at least one*" instead of "*some*". Altogether, 55 items have been constructed – and two instructional/example items. The obligatorily given explanation of logical rules is given in Figure 4.

All lawyers are attorneys. *All judges are lawyers.*	No judge is an attorney. Some judges are attorneys. **All judges are attorneys.** Some judges are no attorneys. All judges are no attorneys

Figure 1
Item 1[55] of the first draft of the test Reality-contradicting Syllogisms (the solution in bold; the first and the fifth answer options have the same meaning. Translation by the authors).

All tits are woodpeckers. *All tits are birds.*	All birds are woodpeckers. No bird is a woodpecker. Some birds are no woodpeckers. All birds are not woodpeckers. **Some birds are woodpeckers.**

Figure 2:
Item 5[55] of the first draft of the test Reality-contradicting Syllogisms (the solution in bold; the second and the fourth answer options have the same meaning. Translation by the authors).

If every Gsels were an Ebira	Every Guatsle would be a Gsels.
and no Guatsle were an Ebira, then...	**No Guatsle would be a Gsels.**
	At least one Guatsle would be a Gsels.
	Some Guatsle would be no Gsels.
	Some Guatsle would be a Gsels.

Figure 3
Item 22[55] of the first draft of the test Reality-contradicting Syllogisms (the solution in bold; the third and the fifth answer options have the same meaning. Translation by the authors).

1. If *all A are B*, then it is not always correct that *all B are A* (e.g.: All poets are humans, but not all humans are poets)
2. If *some A are B*, then it is always correct that *some B are A* (e.g.: Some architects are technicians, therefore some technicians are architects)
3. If *all A are not B*, then it is always correct that *all B are not A* (e.g.: All fish are no reptiles, therefore are all reptiles are no fish)
4. If *some A are not B*, then it is not always correct that *some B are not A* (e.g.: Some dogs are no poodles, however some poodles can be dogs)
5. If *some A are B*, then it is not always correct that *some A are not B* (e.g.: "some" means within logic "at least one", that is theoretically all A can be B)

Figure 4
The used explanation of logical rules (translation by the authors)

When using syllogisms as a task of a psychological test, content validity is likely given with regard to reasoning, the "ability to realize regularities and logically compelling connections in order to put in appropriate use" (Kubinger, 2019, p. 244; translation by the authors). According to the six categories of reasoning tests by Kubinger (2023a) – crossing *fluid* vs. *crystallized* facets with *lexical* vs. *numerical* vs. *figural* contents – the test *Reality-contradicting Syllogisms* meets the combination of *fluid* facets and *lexical* contents. As that combination is hardly available to practitioners, this test's elaboration seems worthwhile: To be logically consequent, unequivocal, and congruent in verbal communication certainly is a skill, required in many professions, for instance for consulting psychologists. Even when the above cited online self-assessment for applicants of *Computer Science & Electrical Engineering* studies was compiled, the corresponding requirements analysis with experts resulted in the need to test reasoning not only (*crystallized*) figural and numerically, but also *fluid-lexically*.

Psychometric analysis with respect to the Rasch model was processed for only 20 out of the 55 items, which could be administered (without any time restriction) to 245 male selectees of the Austrian military service examination. This analysis happened in accordance with the state of the art (cf. Kubinger, 2005). That is, Andersen's Likelihood-ratio test (LRT) was used with several partition criteria of the given overall sample into subsamples (1. score: "high-" *vs.* "low-scorers", meaning the partition in testees with a high number of solved items *vs.* testees with a low number of solved items; 2. level of education (high school): yes *vs.* no; 3. mother tongue: German *vs.*

other than German; 4. provision of aids (paper and pencil) due to an experimental design: yes *vs.* no). In case of any significant LRT (comparison-wise type-I-risk α = .01 – running four comparisons this means a study-wise type-I-risk of approximately α = .04 < α = .05), items were deleted step by step when repeating this model test until it resulted in non-significance for each partition criterion. By means of Rasch's graphical model check, it was decided which item to delete. That check illustrates the congruence of item parameter estimations when based on different subsamples: as a rule of thumb differences of any item parameter estimation between two subsamples larger than a tenth of the parameters' range indicate model misfit (see again Kubinger, 2005).

For analyzing the data, the R-package eRm (Mair & Hatzinger, 2007) was used.

Results

Table 1 summarizes the results of Andersen's Likelihood-ratio test (LRT) with respect to the four partition criteria.

Table 1
The Rasch model tests for 20 items of the first draft of the test Reality-contradicting Syllogisms. For the applied four criteria of partitioning the overall sample, the results of the asymptotically χ^2-distributed Andersen's Likelihood-ratio test statistic (LRT) are given as well as the degrees of freedom (df) and the respective p-value. The results are based on 228 testees (17 proved to work not seriously on the test).

partition criterion	χ^2	*df*	*p*
score	53.13	19	.000
level of education	32.38	19	.028
mother tongue	22.33	19	.268
provision of aids	21.55	19	.307

It results only a single significant LRT, which concerns the partition criterion "high-" *vs.* "low-scorers". The graphical model check in Figure 5 reveals in particular a misfit of item (syl_)8[20]. Given Rasch model's validness, each item has due to "specific objectivity" (cf. Scheiblechner, 2009) the same item (difficulty) parameter, regardless of which subsample is used; as a consequence, opposing the item parameter estimations of two subsamples in a Cartesian coordinate system would ideally result only in dots lying on a 45° line which meets the origin. But actually, item 8's parameter estimation achieves within subsample "low-scores" a much lower value than within subsample "high-scorers". The corresponding confidence ellipse which takes the standard error of item parameter estimation into account, clearly shows the disproportion. This item is in relation to the other items not as difficult for a "low-scorer" as it is for a "high-scorer". Item (syl_)5[20] stands similar out.

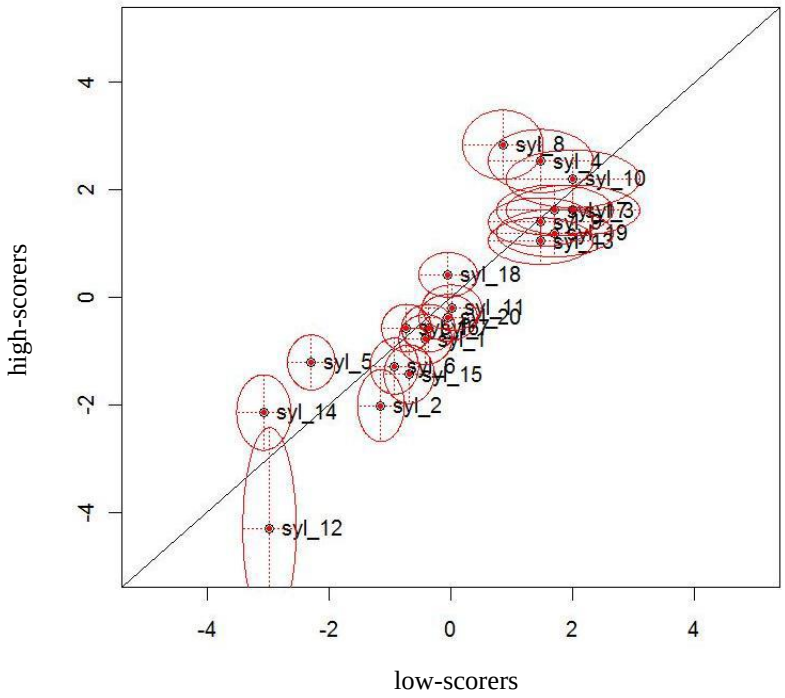

Figure 5
Graphical model check for 20 items of the 55 item pool of the first draft of the test Reality-contradicting Syllogisms – item (difficulty) parameter estimations according to the Rasch model as opposed for selectees with a high score (ordinate) and for selectees with a low score (abscissa). For the items not only the (estimated) item parameters are plotted against each other but the confidence ellipses are also shown. These result when the standard error of item parameter estimation is taken into account (α = .01).

Deleting item 8[20] from the pool and re-analyzing the remaining item pool led again to a significant LRT with respect to the partition criterion "score"; this time, actually item 5[20] was to delete. Again, that item is in relation to the other items not as difficult for a "low-scorer" as it is for a "high-scorer". However, after deleting item 5[20], too, no significant LRT resulted (see Table 2).

Table 2
The Rasch model tests for 18 items of the first draft of the test Reality-contradicting Syllogisms. For the applied four criteria of partitioning the overall sample the results of the asymptotically χ^2-distributed Andersen's Likelihood-ratio test statistic (LRT) are given as well as the degrees of freedom (df) and the respective p-value. The results are based on 228 testees (17 proved to work not seriously on the test).

partition criterion	χ^2	*df*	*p*
score	25.07	17	.093
level of education	31.29	17	.018
mother tongue	21.74	17	.195
provision of aids	22.72	17	.159

On one hand, the results are encouraging, as only two of 20 items have to be deleted in order for the Rasch model to hold *a-posteriori* – according to Kubinger and Draxler (2007) 10 percent of deleted items is the commonly tolerable rate. Supported by the results of Srp (1974) – as mentioned above Rasch model's *a-posteriori* validness was given after only five of 80 items were deleted –, it actually looks like measuring with the syllogisms conceptualization is possible along the Rasch model; this does not only guarantee uni-dimensional measurement but is also essential when (just) the number of solved items shall be scored (see Fischer, 1995, for mathematical proof). On the other hand, the results are disappointing because no explanation can be found why just both these items had to be deleted (see Fig. 6 and 7): Their contents are not obviously different from that of the other items. And formally, both items are of type MP-MS (item 8[22] with the propositions for p1-p2-c: 1-2-2, item 5[20] with the propositions for p1-p2-c: 3-1-4) but so are a number of other items (with either the propositions for p1-p2-c: 1-2-2 or 3-1-4), all of which proved to fit the Rasch model.

Every aliquot is a substitute.	**At least one aliud is a substitute.**
At least one aliquot is an aliud.	At least one aliud is not a substitute.
	No aliud is a substitute.
	Every aliud is a substitute.
	Every aliud is not a substitute.

Figure 6
Item 8[20] of the first draft of the test Reality-contradicting Syllogisms (the solution in bold; the second and the fourth answer options have the same meaning. Translation by the authors).

No quopp is a wabb.	No fif is a wabb.
Every quopp is a fif.	All fifs are wabbs.
	Some fifs are not wabbs.
	Every fif is not a wabb.
	Some fifs are wabbs.

Figure 7
Item 5[20] of the first draft of the test Reality-contradicting Syllogisms (the solution in bold; the second and the fourth answer options have the same meaning. Translation by the authors).

The item parameters for the remaining here analyzed 18 items lie between -2.38 and 3.27. This range of item parameters sounds from experience of a medium extent.

Discussion

The conceptualization of (*Reality-contradicting*) *Syllogisms* stood again the test as concerns the validness of the Rasch model. However, both the here deleted items as well as those which were deleted by Srp (1994) indicate in no way any reason why they do not fit the model. For now, these items' non-conformity with the Rasch model is simply be attributed to chance. But for deeper understanding of the challenge of syllogisms as a psychological test's task further research is needed. The items of a next draft should submit first to the method of thinking aloud, in order to detect early enough unexpected formulation or arrangement problems of some acting terms or distractors; and a statistical distractor analysis would above all disclose which distractor is marked striking frequently and therefore indicates a systematically misleading task. Furthermore, knowledge is needed on whether there are initial learning effects and if so, when, i.e. after how many items, such learning effects will be commonly completed. That might be tested by applying several different sequences of item presentation, allowing a comparison of the difficulty parameters of the same item, presented once very soon, the other time very late. Maybe there are even type- and/or proposition-specific learning effects. If any learning effect takes longer than five items, the test *Reality-contradicting Syllogisms* turns out not to be a "status test", but rather a "learning test" *sensu Guthke* (cf. Guthke, 1992) – meaning that uni-dimensional measurement according to the Rasch model does not work. Otherwise, up to five warming-up items would be needed in addition to the two instructional/example items, which are already given.

According to experience in the administration of this type of task for demonstration purposes in university education, many people have a particular aversion to such a test. This is due to the big challenge of applying logical rules appropriately, above all when the acting terms are set in reality-contradicting relations. That is, such a test

means a very high energetic-motivational demanding. Thus, calibrating the items of the next draft using data from testees for whom nothing depends on the test score is probably even less appropriate than for other psychological tests. If volunteers abandon their achievement motivation sooner or later, several items towards the end of the test would result disproportionately more difficult than the items presented earlier. Therefore, even more than in other cases, the testees used for item calibration should come from a population for which the test result has consequences – as happened in the described study.

Whether *fluid-lexical* reasoning as aimed to be measured with the test *Reality-contradicting Syllogisms* constitutes indeed a specific intelligence factor but is not covered by other factors or even a general reasoning factor, is for the time being examined by Kubinger (2023b).

References

Fischer, G. H. (1995). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models* (pp. 15–38). New York: Springer.

Freytag-Löringhoff, B. (1961*). Logik – Ihr System und ihr Verhältnis zur Logistik* [The system of logic and its relation to logistic]. Stuttgart: Kohlhammer.

Guthke, J. (1992). Learning tests-the concept, main research findings, problems and trends. *Learning and Individual Differences, 4*, 137-151.

Kubinger, K. D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*, 377-394.

Kubinger, K. D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (eds.), *Principles and methods of test construction* (pp. 104-119). Göttingen: Hogrefe.

Kubinger, K. D. (2019). *Psychologische Diagnostik – Theorie und Praxis psychologischen Diagnostizierens* (3rs ed.) [*Psychological Assessment – Theory and Practice of Psychological Consulting*]. Göttingen: Hogrefe.

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

Kubinger, K. D. (2023a). Guest Editorial: Promising reasoning test ideas not yet published. Special issue: *Promising reasoning test ideas not yet published. Psychological Test and Assessment Modeling, 65*, 315-320.

Kubinger, K. D. (2023b). On the dimensionality of Reasoning. *Psychological Test and Assessment Modeling, 65*, 437-447.

Kubinger, K. D. & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell [Problematic issues when calibrating a psychological test according to the Rasch model]. *Diagnostica, 53*, 131-143.

Mair, P. &, Hatzinger, R. (2007). eRm: Extended Rasch Modeling. 0.9-5. http://cran.r-project.org/web/packages/eRm/eRm.pdf/

Treiber, S. (2012). Schlussfolgerndes Denken lexikalisch: Der Wiener Syllogismentest [Lexical reasoning: The viennese syllogisms test]. In K. D. Kubinger, M. Frebort, L. Khorramdel, & L. Weitensfelder („Wiener Autorenkollektiv Studienberatungstests“ [“Viennese Authors Collective“]) (eds.), *Self-Assessment: Theorie und Konzepte* [*Self-Assessment: Theory and conceptualizations*] (pp. 177-180). Lengerich: Pabst.

Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. *Psychology Science Quarterly, 51,* 181-194.

Appendix

The 19 conclusive syllogisms (two premises and the conclusion; the latter in bold)

All M are P. All S are M. **All S are P**.	All M are P. All M are S. **Some S are P**.	All M are P. Some M are S. **Some S are P**.	All M are P. Some S are M. **Some S are P**.
Some M are P. All M are S. **Some S are P.**	Some P are M. All M are S. **Some S are P**.	All P are M. All S are not M. **All S are not P**.	All P are M. All M are not S. **All S are not P**.
All M are not P. All S are M. **All S are not P**.	All P are not M. All S are M. **All S are not P**.	All M are not P. All M are S. **Some S are not P**.	All P are not M. All M are S. **Some S are not P**.
All M are not P. Some M are S. **Some S are not P**.	All P are not M. Some M are S. **Some S are not P**.	All M are not P. Some S are M. **Some S are not P**.	All P are not M. Some S are M. **Some S are not P**.
Some M are not P. All M are S. **Some S are not P**.	All P are M. Some S are not M. **Some S are not P**.	All P are M. All M are S. **Some S are P**.	