LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Georg Schollmeyer, Christoph Jansen, Thomas Augustin

A simple descriptive method for multidimensional item response theory based on stochastic dominance

Technical Report Number 210, 2017
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

A simple descriptive method for multidimensional item response theory based on stochastic dominance

Georg Schollmeyer Christoph Jansen Thomas Augustin

Abstract

In this paper we develop a descriptive concept of a (partially) ordinal joint scaling of items and persons in the context of (dichotomous) item response analysis. The developed method has to be understood as a purely descriptive method describing relations among the data observed in a given item response data set, it is not intended to directly measure some presumed underlying latent traits. We establish a hierarchy of pairs of item difficulty and person ability orderings that empirically support each other. The ordering principles we use for the construction are essentially related to the concept of first order stochastic dominance. Our method is able to avoid a paradoxical result of multidimensional item response theory models described in Hooker et al. [2009]. We introduce our concepts in the language of formal concept analysis. This is due to the fact that our method has some similarities with formal concept analysis and knowledge space theory: Both our methods as well as descriptive techniques used in knowledge space theory (concretely, item tree analysis) could be seen as two different stochastic generalizations of formal implications from formal concept analysis.

Keywords: stochastic dominance, empirically mutually supportive pairs, formal conept analysis, knowledge space theory, cognitive diagnosis models, formal implications, item tree analysis

1 Introduction

The fruitful development of psychometric models that are empirically adequate is always challenged by the fact that one has to deal with latent constructs that, without any quasi metaphysical theory or vague preconceptions of involved terms, lead to a systematic under-determination of involved concepts. A sound scientific theory that makes all basic terms like that of ability or difficulty rigorously empirically criticizable is difficult to obtain. The exact understanding of the latent structure as a truly underlying trait or only as a merely rough sketch or simply only an analogy about what is going on behind the observable scene has a straight impact on how to interpret results of psychometric tests and how to deal with a seemingly paradoxical situation, firstly described in Hooker et al. [2009]. This paradox is prevalent in many multidimensional models of item response theory (IRT), including Rasch-type models that are not only statistical models for analyzing item response data, but have some seemingly solid measurement theoretic foundation that can possibly be shattered by such a paradox.

The aim of the present paper is to provide a more or less naive descriptive viewpoint on the problem of obtaining some notion of person ability and item difficulty, given one simply

has observed, how a set of persons answered a set of items in a psychometric test. Our descriptive method avoids the above-mentioned paradox.

The paper is structured as follows: In Section 2 we recall the paradox discovered by Hooker et al. [2009] and give a possible explanation of the paradox. In Section 3 we introduce our descriptive method for a relational notion of person ability and item difficulty. We describe our ideas in the language of formal concept analysis. The basics of formal concept analysis are briefly sketched in Appendix A for the reader unfamiliar with this topic. Actually, for the understanding of Section 3 it is not really needed to know much about formal concept analysis. The reason for presenting our ideas in the language of formal concept analysis is that it has a neat relation to knowledge space theory (and also to cognitive diagnosis models), which is a nonparametric form of item response theory where the analysis of the paradox is also of some interest. In Section 4 we analyze the paradox in the context of knowledge space theory. A brief introduction to the basics knowledge space theory is also given in Appendix B. Appendix C shortly sketches the relationship between formal context analysis and knowledge space theory. Section 5 gives a brief data example comparing the herein developed method with a descriptive method of item tree analysis and sketches, how our method behaves under certain unidimensional item response models. Finally, Section 6 concludes.

2 The paradox

We start by explaining the paradox we are referring to:

> *"Jane and Jill are fast friends who are nonetheless intensely competitive. At the end of high school, they each take an entrance exam for a prestigious university. After the exam, they compare notes and discover that they gave the same answers for every question but the last. On checking their materials, it is clear that Jane answered this question correctly, but Jill answered incorrectly. They are therefore very surprised, when the test results are published, to find that Jill passed but Jane did not.*
>
> *Lawsuits ensue. The university maintains that it followed well-established statistical procedures: The questions on the test were designed to simultaneously examine both language and analytic skills, and a multiple-hurdle rule (Segall, 2000) based on maximum likelihood estimates of each student's abilities was used to ensure that admitted students were proficient in both. The university had rechecked its calculations many times and was satisfied the correct decision had been made. Jane's lawyers countered that, whatever the statistical correctness of the agency's procedures, it is unreasonable that an examinee should be penalized for getting more questions correct."* ([Hooker et al., 2009, p. 419])

Before coming to an explanation and discussion of this paradox we want to emphasize the fact that the paradox is not a marginal note arising in only a few instances of multidimensional IRT models and data situations, it can be shown that this sort of paradox can (and to some extent will) arise in broad classes of multidimensional IRT models (cf. Hooker et al. [2009], Jordan and Spiess [2012], Finkelman et al. [2010]), it is also present in some models of knowledge space theory (see Section 4). Note further that the paradox does not arise only

because of the need to estimate an underlying ability of a person, which is subject to statistical error. Also if one would replace the actually observed responses patterns by true underlying response probabilities one could get such a paradox. (Of course, for many usual models, such paradoxical situations are somehow asymptotically avoided, given the presumed model is actually true, cf., [Jordan, 2013, Section 4.1] and [Hooker et al., 2009, Theorem 8.1]). This is important to keep in mind to understand the next considerations. An intuitive explanation why the paradox can appear is given in Hooker et al. [2009]:

> *"Suppose that each question on the test given to Jane and Jill required both language and analytical skills to answer correctly, and Jane and Jill got some of these correct and some incorrect. The final question, however, was very difficult in terms of analysis, but did not require strong language skills. That Jane got this question correct suggests that her analytical skills must be very good indeed. This being the case, the only explanation for her previous incorrect answers is that her language skills must be quite low. By contrast, Jill, in getting the final question incorrect, has demonstrated fewer analytic skills and must have relied on stronger language skills to answer previous questions correctly. The estimate of Jane's language ability therefore dipped below the required threshold, while Jill's was pushed upward; both obtained satisfactory analysis scores."* ([Hooker et al., 2009, p. 420])

The authors of Hooker et al. [2009] conclude that they *"nonetheless feel that it is better not to put students in the position of second-guessing when their best answer may be harmful to them"* ([Hooker et al., 2009, p. 420]), however, there are different reactions thinkable. Three of many possible reactions are:

i) **"Everything is alright** here, because given that the model is true, we have appropriately estimated the true abilities of Jill and Jane and thus appropriately decided for Jill because of her sufficient language skills and against Jane because her language skills were not sufficient.

ii) **Maybe everything is alright, but** at least from the point of view of some notion of "fairness" (which is a concept independent of the notions of ability and difficulty) we should only use psychometric methods for the approval of students that do not admit this paradox.

iii) **The paradox reveals two very different conceptualizations** of the term ability that should not be confused: On the one hand one can understand the term ability as an underlying trait that one tries to measure, and thus the responses given in the exam have to be taken only as measurements that are only indirectly related to the underlying trait. On the other hand one can understand the ability literally as the ability to solve this or that item and so one has to take the responses not only as measurements but as the actual results showing exactly which items a persons was able to solve in the concrete test and which not. In this understanding, one could alternatively use the word success instead of the word ability. Then there would be no need in estimating abilities, but a naturally arising question would then be, which of two persons that solved different questions was more successful.

The first reaction is problematic in the sense that one cannot assume that the model is exactly right, because a rigorous empirical test of such an IRT model is not really possible due to the involved latent concepts (see also the discussion in [Michell, 2008b]). Especially the dimension of the IRT model plays an important role, here: If we could have (and one should have) doubt about the (clearly only very roughly adequate) statement "the test items only test the dimensions *language skill* and *analytical skill*", then the situation may change: In a very extreme (and of course unrealistic) opposite situation one could assume that in case of doubt every item tests mainly its own dimension. Then, clearly the paradox should not be present anymore, because then Jane would be more able w.r.t. every dimension that was tested.

The second reaction is maybe confronted with the objection "Why is it unfair to base an approval on honest estimates of abilities and not to revise a decision only because of the fact that actually Jane did answer one more question rightly? The fact that Jane answered one more question rightly is only revealing that she solved the items actually given in the test better and not that she has more abilities." If one accepts the second reaction, then one could do for example a constrained optimization, like proposed in Hooker et al. [2009].

The third reaction is of interest in this paper, where we take the response patterns at face value to define a descriptive notion of success.

3 A descriptive method based on concepts of stochastic dominance

In this section, we develop a purely descriptive and relational notion of person success and item difficulty. We start with a motivating example to introduce our ideas:

3.1 A motivational example

> Tim and Danny are companioned pole vaulters. From time to time they discuss about the adequacy of the pole vault rules which actually made Danny better ranked than Tim in the last three competitions, seemingly only because he was more smart in skipping the right heights: In 2015 everything was in order, Tim passed the $5.70m$ in the first trial and failed the following $5.75m$ while Danny passed both heights in the first trail. In 2016 their difference was more tight: While both failed the $5.70m$ for the first two trials, Tim took the $5.70m$ in the third trial but Danny skipped the $5.70m$ and passed the $5.75m$ in the first trial. The height of $5.80m$ they both did not manage. Finally, in 2017 Danny's luck in skipping heights beggared all description: Again Tim passed the $5.70m$ in the third trial and then took all heights till $5.90m$ in the first trial until he failed the $5.95m$. But Danny, also passing the $5.70m$ in the third trial, decided to skip the following 4 heights and luckily passed the $5.95m$ in the first trial.
>
> In discussions between Danny and Tim, Danny usually argues that the rules are simply the rules and thus skipping heights is of course right if the luck is with you. But Tim objects:

"Beside the rules, what would be a reasonable argumentation showing that in the last three years you were actually the more capable pole vaulter?" Danny interrupts:
"If you had force me to also jump the skipped heights between $5.75m$ *and* $5.90m$ *then I would probably have passed it, because I also managed the* $5.95m$*."*
Tim: *"Oh, this is wild speculation. But let us make the problem more simple: Assume we both did jump the same heights, say only one time, and you failed all heights from* $5.75m$ *to* $5.90m$ *while I passed all these heights. Would you still say that you are the "better" pole vaulter only because you managed the* $5.95m$ *and I not?"*
Danny: *"Of course I would say this,* $5.95m$ *is an extraordinary performance, near to the olympic record."*
Tim: *"Oh, I doubt that the difference between* $5.90m$ *and* $5.95m$ *is so much to make up for four failed* $5.90m's$*."*
Danny: *"You cannot compare one passed* $5.95m$ *with* 4 *failed heights of* $5.90m$*."*

Tim: *"Of course I cannot compare, but also you cannot compare, so what could we do?..."*

The situation of pole vaulting has some similarities with the situation in item response theory, both Tim and Danny are solving items, but there are important differences, which make the analysis more simple, here. Concretely, we have the following two specific points:

i) The items of jumping a specific height have some "intrinsic" properties that make them totally ordered in difficulty. If one idealizes the situation a little bit, then one can say that if one is able to take some height c, then one is also able to take a height d that is lower than c. To see this, one can argue that for jumping the height d, one can put the bar at height d and jump, as if the bar would lie on height c, and because one is able to jump height d, one would automatically also take height c. It is important to note that this relational property between height c and height d is an "intrinsic" property of the items "jumping height c" and "jumping height d", which is due to some physical relation between the tasks. In particular, it is not dependent upon who is trying to solve the task, so here we have no dependency of the notion of "difficulty" upon a population. This independence is also the aim in item response theory in Rasch's understanding (*specific objectivity*, cf., [Rasch, 1977]), but there it often seems to be more a wish than a fact.

ii) The task of jumping a given height can be repeated by the same person. The repetition of the trials of one fixed person jumping a fixed height can be idealized as a binomial experiment where one repeats a number of Bernoulli trials with some success probability p. Approximately, the trials can be assumed to be independent, at least if there is enough time for regeneration between different trials and if possible learning effects can be neglected. In typical situations of item response theory, it is not useful to pose the same question twice to the same person because of the presence of strong learning effects. The fact that the pole vaulting trials can be repeated makes a statement like "Tim has a probability of taking $5.90m$ of around 0.9" a somehow empirically testable statement[1],

[1] At least if one accepts Poppers *methodological decision* for a "practical falsification", cf., [Popper, 2005, p. 182]

while in typical situations of item response theory, one cannot test a comparable statement for one person and one item in isolation. One would have to rely on either other items, from which one knows (but wherefrom?) that they have the same difficulty, or on items, from which one can somehow translate the probabilities of solving that items to the probability of solving the envisaged item. For this one needs a model like the Rasch model that has to be also valid to allow for such a translation. But beforehand one does not know that the Rasch model holds. In this sense, in the spirit of the Duheme-Quine problem one can test one single item and one single person only together with e.g. the Rasch model.

3.2 Person ability and item difficulty in pole vaulting

So, let us now try to motivate some notion of item difficulty and person ability firstly for the simple situation of pole vaulting. As already said, there is some clear total order between the different items of jumping given heights that are due to physical relations. One can ask if there is more than an ordinal scale, for example a cardinal scale of measurement underlying, here. The heights itself clearly have a cardinal scale of measurement. But in pole vaulting, one is not interested mainly in the heights itself, but in the question, if one is able to take, or if one actually did take a given height. If one thinks in the probability of taking a given height, then it is reasonable to assume that the success probability is decreasing in the height, but there seems to be no obvious specific functional form describing the success probability in dependence on the height. Actually, the reason for sometimes managing a given height and sometimes not is due to auxiliary conditions that are not genuinely related to the height or the experiment, but to the not explicitly stated circumstances.

Thus, practically, more than an ordinal scale of measurement seems to be not reachable, here. Of course, in principle, if we exactly know the auxiliary conditions then we possibly can explicitly determine the quantitative relation between the success probabilities and the height and other auxiliary conditions. Practically, this seems unrealistic.
Compared to the auxiliary conditions that introduced the randomness of sometimes taking a height and sometimes not, in the Rasch model, the randomness in solving an item is not a sort of noise from which one would like to get rid of, instead it is the basic ingredient that allows for the construction of a cardinal scale of item-difficulty. If no randomness were present, then strangely enough one would fall back into the situation of Guttman scaling, where one has only an ordinal scale of measurement. This counterintuitive issue is known as the Rasch paradox (see, [Michell, 2008a,b], cf., also [Sijtsma, 2012, Humphry, 2013]).

Since we are not concerned with more than a (partially) ordinal scale in this paper, we do not have to care about this issue, here.

To get a notion of ability of Tim and Danny seems to be hard. For the item difficulty, we got only a relational notion of difficulty, thus we would also only expect a relational notion for person ability. We said that some height c with $c > d$ is more difficult than a height of d because if one is able to jump the height c, then one is also able to jump the height d. In a dual manner one can say that one person p is more able than another person q, if person p is able to take all heights that person q is able to take. Here, we face the first slight asymmetry between person ability and item difficulty[2]. For the items, we have some physical relation

[2]Note that a further asymmetry consists in the fact that the ability of a person is related to which item one

that translates to a relation about the difficulty of items. If we could analyze both Tim and Danny as two different "physical machines" then we could possibly get also a notion of which heights Danny and Tim are physically able to take. This would be fine, but is of course too difficult. A further point is here, that we spoke about the ability to solve an item, but of course we actually meant some notion of "being **in principle** able to take a height", because sometimes one takes a height and sometimes not. For the comparison of difficulty we did not need to rely on which heights Danny and Tim actually did take. For the comparing of persons, it seems to be practically not circumventable to rely on which items the persons actually did solve.

3.3 A notion of item difficulty and person success in item response theory

To get a notion of person ability, let us firstly think about the still more simple case where one has items that are completely comparable, say Danny and Tim did only try to take the same height c for a number of trials. Then one can naturally say that a person who solved more items than another person is more able than the other person. Actually, because of the randomness of solving an item or not, we would like to be a little more cautious here and say only that the person who solved more items is more successful than the other person. The above notion can be mathematically expressed in different ways. One way would be to calculate the item scores (i.e., the number of solved items) and then to say that the person with a higher score is more successful. This formulation seems to do not naturally translate to the case where we have items with different difficulties with an only ordinal scale of measurement. Thus, another representation seems to be promising, here:

For the case of items that are clearly comparable in difficulty, one can say that person p is more successful than person q, if for every item i that person q did solve, there exists also an item $\Phi(i)$ that person p did solve, with the additional condition that Φ is injective meaning that we use no item from the items that p has solved twice as an argument to show that person p is more successful than person q. This representation with a sort of a matching is the main idea in this paper, and this idea is closely related to the notion of stochastic dominance, see Appendix D. Now, we can think about how we would generalize this notion to the case of items with different difficulties. The generalization is actually very intuitive: For two persons p and q one can naturally define that person p is more successful than person q if for every item i there (bijectively) exists another item $\Phi(i)$ that is as least as difficult as item i and was solved by person p.

After having found a notion of successfulness for persons if the difficulties of items are given, we can now think about how to define a successfulness-relation when there is no difficulty relation given beforehand, which is the typical case in item response theory.

Before doing so, we would firstly like to point out that there are still very interesting situations, in which one has at least a partial ordering of item difficulty beforehand that is due to an understanding of the cognitive processes that are needed to solve an item. A classical example are fraction subtraction tasks in the context of cognitive diagnosis models

considers (e.g. pole vaulting or climbing or whatever) whereas the difficulty of an item is more intrinsically related to the item and not so much to the question about who tried to solve it.

(CDM: Bolt [2007], de la Torre [2009], DiBello and Stout [2007], Junker and Sijtsma [2001], Tatsuoka [1990, 2002]). If one compares for example solving the subtraction task $6/7 - 4/7$ with the task $2 - 1/3$, then one can say that the first task is easier than the second, because for solving the first task one has essentially only to subtract the nominators, while for the second task, one has to firstly convert the 2 to $6/3$ and then one has to subtract the nominators. Thus, everyone who can solve the second task usually is also able to solve the first task, and in this sense the first task is easier, which is a relation inherent in the tasks and not related to the person who solves the task. Apart from cognitive diagnosis models, also in knowledge space theory ([Doignon and Falmagne, 1985, 2012, Falmagne et al., 1990, Falmagne and Doignon, 2010]) one similarly has some relations between different items. Actually, there is a neat relation between the two areas which seemingly independently developed very similar concepts. While in knowledge space theory one firstly had a more deterministic view that was generalized to probabilistic versions afterwards (e.g., the basic local independence model.), in cognitive diagnosis modeling one came more from classical item response theory with a clear probabilistic underpinning and had the aim of bringing item response theory closer to cognitive psychology. For a detailed discussion of the link between cognitive diagnosis models and knowledge space theory, see Heller et al. [2015]. Note further, that both knowledge space theory and cognitive diagnosis models are somehow related to the theory of formal concept analysis, see Rusch and Wille [1996]. This relation is the reason for presenting our ideas in terms of formal concept analysis. (Actually, for the understanding of the ideas developed herein, one does not need to know much about formal concept analysis, basically one only needs to know, what a formal context is and what a formal implication is, see Appendix A). The paradox described in Hooker et al. [2009] was given in the context of more classical multidimensional IRT models, but one can also ask for the presence of the paradox in the context of knowledge space theory (or cognitive diagnosis models). This question will thus be also discussed in Section 4 of this paper.

Coming back to the problem of defining a successfulness relation in a general IRT situation where one has neither some "physical insights" into person abilities nor a relation on the item difficulties, at first glance it seems to be impossible to get a reasonable successfulness relation. However, in a first step, one can try to solve the problem of a missing item difficulty relation by thinking about a dual construction that could lead to a reasonable difficulty relation in a situation where one has knowledge about the abilities of the persons. The following notion of item difficulty, given a notion person ability seems to be natural:

For two items i and j and for a set of persons with the same ability who all tried to solve item i and item j one can say that item i is more difficult than item j if there are less persons who solved item i than persons who solved item j. Equivalently, one can say that item i is more difficult than item j if for every person p who solved item i there bijectively exists a person $\Psi(p)$ who solved item j. For persons with different abilities, one can say that item i is more difficult than item j if for every person p who solved item i there bijectively exists a person $\Psi(p)$ who is not more able than person p and solved item j.

Now, of course one does not know beforehand the abilities of the persons and one is thus running in circles. However, this running in circles can be made from a necessity to a virtue by simultaneously treating item difficulty and person success relations. A difficulty relation induces a successfulness relation and vice versa, so one can think about pairs of relations that fit

to each other. This is the idea behind the notion of empirically mutually supportive pairs that we would like to introduce more formally in the following section. The notion of empirically mutually supportive pairs will be closely related to the notion of first order stochastic dominance, see Appendix D. Since there are more equivalent definitions of stochastic dominance, we also give two equivalent and intuitively accessible definitions of associated success and easiness relations in Definition 1 which prepares the notion of empirically mutually supportive pairs given in Definition 2.

3.4 Empirically mutually supportive pairs

Definition 1. *Let $\mathbb{K} := (G, M, I)$ be a formal context which means that G and M are sets and $I \subseteq G \times M$ is a binary relation between G and M. In our context the set G is the set of persons participating in a dichotomous psychometric test consisting of the set M of items. A pair (g, m) is in I if and only if person g has solved item m.*

For a given person g we denote with g' the set of all items that person g has solved. Analogously for an item m we denote with m' the set of all persons who solved item m. Let furthermore $S \subseteq G \times G$ be a quasiorder on G and let $E \subseteq M \times M$ be a quasiorder on M. The quasiorder S is interpreted as: gSh means that person g is "more successful" than person h according to the relation S. The quasiorder E is interpreted as: mEn means that item m is "easier" to solve than item n according to the relation E. To the quasiorders S and E we associate the following two relations:

$$\mathbb{SD}(E) := \{(g,h) \in G \times G \mid \underbrace{\forall U \in \mathcal{U}((M,E)) : |U \cap g'| \geq |U \cap h'|}\ \}$$

Person g ist more successfull than person h, if she solved more (or as many) "non-easy" items than (as) person h, independtly from the exact concretization of the term non-easy with the help of upsets of (M, E) or equivalently.:
Person g is more successfull than person h, if for every item solved by person h there bijectively exists an item at least as difficult that was solved by person g

and

$$\mathbb{SD}(S) := \{(m,n) \in M \times M \mid \underbrace{\forall U \in \mathcal{U}((G,S)) : |U \cap m'| \geq |U \cap n'|}\ \}$$

Item m is easier (or as easy) to solve than (as) item n, if it was solved by more "less successfull" persons than item n, independently from the exact concretization of the term less successfull with the help of upsets in (G, S) or equivalently:
Item m ist easier (or as easy to solve) than (as) item n, if for every person that solved item n there bijectively exists a less or equally successfull person who solved item m.

Here, $\mathcal{U}((M, E)$ denotes the set of all upsets of the quasiorder (M, E), where a set $S \subseteq M$ is called an upset if it satisfies $a \in S$ & $aEb \Longrightarrow b \in S$. One can understand the set of all upsets as the set of all reasonable concretions of the term non-easy: an upset A is a concretion of the term non-easy by declaring all $a \in A$ as non-easy and all $a \notin A$ as easy. The concretion is reasonable in the sense that if a is termed non-easy and a is easier than b (i.e. aEb), then also b should be termed non-easy, which is exactly the property characterizing an upset. Analogously, $\mathcal{U}((G, S))$ defines the upsets of the quasiorder (G, S) modeling the set

of all reasonable concretions of the term non-successful.

Now, define additionally the space $\mathfrak{S} := \{(S,E) \mid S \in 2^{G\times G}, E \in 2^{M\times M}\}$ *and endow* $\mathfrak{S}$ *with the order*

$$\leq_{\mathfrak{S}} := \{((S,E),(S',E')) \in \mathfrak{S}\times\mathfrak{S} \mid S \subseteq S' \,\&\, E \subseteq E'\},$$

and finally define the operator

$$\mathfrak{L} : (\mathfrak{S},\leq_{\mathfrak{S}}) \longrightarrow (\mathfrak{S},\leq_{\mathfrak{S}}) : (S,E) \mapsto \mathfrak{L}((S,E)) := (\mathbb{SD}(E)\cap S, \mathbb{SD}(S)\cap E).$$

Definition 2. *Let* $\mathbb{K} := (G,M,I)$ *be a formal context and* $(S,E) \in \mathfrak{S}$ *a pair of a person- and an item-quasiorder. The pair* (S,E) *is called a* **weakly empirically mutually supportive pair of** $\mathbb{K}$*, if*

$$S \subseteq \mathbb{SD}(E) \quad \& \quad E \subseteq \mathbb{SD}(S)$$

or equivalently

$$(S,E) = \mathfrak{L}((S,E))$$

holds. If we actually have

$$S = \mathbb{SD}(E) \quad \& \quad E = \mathbb{SD}(S),$$

then the pair (S,E) *is called a* **strongly empirically mutually supportive pair of** $\mathbb{K}$*. The set of all weakly empirically mutually supportive pairs of* $\mathbb{K}$ *is denoted with* $wsupp(\mathbb{K})$ *and the set of all strongly empirically mutually supportive pairs of* $\mathbb{K}$ *is denoted with* $ssupp(\mathbb{K})$ *.*

We can interpret a given weakly empirically mutually supportive pair (S,E) as follows: If we would have some reason to believe that E is the "truly underlying" easiness relation of the items, then the relation S would be a reasonable relation that is somehow empirically supported by the easiness relation E. Dually, if we have some reasons to think that S is the truly underlying success relation, then E would be empirically supported by S as a reasonable easiness relation. So,we can think of mutually supportive pairs as **possible** underlying pairs of relations that are consistent in the sense that they support each other. There are much of these pairs and one can think about the structure of these pairs. It will turn out that there is a weakest and a strongest such pair. More importantly, one can explicitly compute these both extreme pairs. Furthermore, in some sense, all empirically mutually supportive pairs will avoid the paradox of Hooker et al. [2009] in the sense that a person, who solved all items another person solved, plus some more, is always more successful w.r.t. the relation S than the other person. Actually, the weakest mutually supportive pair (S,E) exactly given by

$$pSq \iff \text{person } p \text{ solved all items that where solved by person } q$$

and

$$iEj \iff \text{item } i \text{ was solved by all persons that solved item } j.$$

The still more interesting pair is the strongest pair that could be seen as the strongest relational notion of difficulty and success one could expect to get, only based on the data. To see, that there actually exists such a strongest pair and to see how to compute it, we only have to analyze, what happens if we apply the operator $\mathfrak{L}$ several times.

Lemma 1 (Lemma and Definition)**.** *Let* $\mathbb{K} = (G,M,I)$ *be a finite formal context (meaning that* G *and* M*, and thus also* I *are finite). The operator* $\mathfrak{L}$ *has the following properties:*

i) it is monotone: $\forall p, q \in \mathfrak{S} : p \leq_{\mathfrak{S}} q \Longrightarrow \mathfrak{L}(p) \leq_{\mathfrak{S}} .\mathfrak{L}(q)$

ii) $\mathfrak{L}$ *is intensive:* $\forall p \in \mathfrak{S} : \mathfrak{L}(p) \leq_{\mathfrak{S}} p$.

iii) $\mathfrak{L}$ *is of finite order:* $\exists k \in \mathbb{N} : \mathfrak{L}^{k+1} := \underbrace{\mathfrak{L} \circ \mathfrak{L} \circ \ldots \circ \mathfrak{L}}_{k+1 \text{ times}} = \mathfrak{L}^k$.

iv) If we define $\mathfrak{L}_\infty$ *as* $\mathfrak{L}^k$ *with the* k *from iii), then* $\mathfrak{L}_\infty$ *is a kernel operator, that is, a monotone, intensive and idempotent operator, where idempotent means that* $\mathfrak{L}_\infty^2 = \mathfrak{L}_\infty$.

Proposition 1. *Let* $\mathbb{K} = (G, M, I)$ *be a finite formal context.*

i) The set wsupp$(\mathbb{K})$ *of all weakly empirically mutually supportive pairs of* $\mathbb{K}$ *are exactly the kernels of the kernel operator* $\mathfrak{L}_\infty$. *(This means that* $p \in wsupp(\mathbb{K}) \iff \mathfrak{L}_\infty(p) = p$*).*

ii) The set ssupp$(\mathbb{K})$ *of all strongly empirically mutually supportive pairs of* $\mathbb{K}$ *is a subset of* $\mathfrak{S}$ *that has a smallest and a greatest element.*

The smallest element $(\mathcal{I}^\partial_{G,1}(\mathbb{K}), \mathcal{I}^\partial_{M,1}(\mathbb{K}))$ *of this interval consists of the dual relation of the simple formal implications between objects*

$$\mathcal{I}^\partial_{G,1}(\mathbb{K}) := \{(h, g) \in G \times G \mid \forall m \in M : gIm \Longrightarrow hIm\}.$$

and of the dual relation of the simple formal implications between attributes

$$\mathcal{I}^\partial_{M,1}(\mathbb{K}) := \{(n, m) \in M \times M \mid \forall g \in G : gIm \Longrightarrow gIn\}.$$

The greatest element is given as

$$\mathfrak{L}_\infty((G \times G, M \times M))$$

or equivalently as

$$\mathfrak{L}_\infty((\#_G, \#_M)),$$

where $\#_G := \{(g, h) \mid |g'| \geq |h'|\}$ *and* $\#_M := \{(m, n) \mid |m'| \geq |n'|\}$.

4 Relation to knowledge space theory and formal concept analysis

The reason for introducing our ideas in the language of formal concept analysis lies in the fact that the weakest empirically mutually supportive pair is build by simple formal implications and that the construction of success-relations from easiness-relations and vice versa can be seen as some stochastic generalization of simple implications based on ideas of stochastic dominance. In knowledge space theory, which in its deterministic form is closely related to formal concept analysis (see [Rusch and Wille, 1996] and Appendix C), for the construction of the knowledge structure one sometimes uses techniques of Boolean analysis, for example item tree analysis (cf., e.g., [Schrepp, 1999, 2002, Ünlü and Sargin, 2010]). This descriptive technique can be also seen as some other type of a generalization of simple formal implications:
If in an IRT data set, all persons, who solved item i did also solve item j, then the formal implication $i \longrightarrow j$ is valid and one would naturally say that item j seems to be more easy than

item i. For our stochastic generalization, if for simplicity all persons have the same ability, we would also say that item j is more easy than item i if only more persons solved item j than item i, no matter, which persons exactly solved the items. Another generalization of the formal implication $i \longrightarrow j$ would be to say that item j is more easy than item i if the implication $i \longrightarrow j$ is not exactly true but approximately in the sense that from the population of persons who solved item i most persons (say more than $c \cdot 100\%$ of this population) did also solve item j. This notion, which is very close to the notion of statistical preference[3] ([De Schuymer et al., 2003b,a]), is one of the underlying ideas in descriptive methods of item tree analysis.

The links between our stochastic generalization of simple implications and knowledge space theory also motivate the question if the paradox from above can also appear in applications of knowledge space theory. This section thus analyzes, in which situations the paradox can occur. Firstly, because in knowledge space theory one has no ability parameters but only knowledge states, we have to define what it means that the paradox occurs. Concretely, we deal here only with the simplest probabilistic version of knowledge space theory, namely with the basic local independence model (BLIM, see below). There, one has some observed response patterns and tries to estimate the true underlying knowledge states and the paradox translates into the question about if it is possible that a person p who solved all items another person q solved, plus some more, gets an estimated knowledge state that is not a superset of the estimated knowledge space of person q.

Definition 3 (Presence of the paradox in knowledge space theory)**.** *Let $(\mathcal{Q}, \mathcal{K})$ be a knowledge space[4] and let $f : 2^{\mathcal{Q}} \longrightarrow \mathcal{K}$ be an estimator that maps every observed response pattern to an estimated knowledge space. Then we say that the paradox is present for the two response patterns R and S if*

$$R \subseteq S \text{ but } f(R) \not\subseteq f(S).$$

Since we sometimes have to deal with ties in the sense that for example for maximum likelihood estimators the argmax could be non-unique we will say analogously for a set-valued estimator $f : 2^{\mathcal{Q}} :\longrightarrow 2^{\mathcal{K}}$ and two response patterns R and S that the paradox is present if $R \subseteq S$ and $f(R) = \{T\}$ and $f(S) = \{U\}$ with $T \not\subseteq U$.

Remark 1. *In Hooker et al. [2009] the paradox was defined for a score that would translate in our context to a mapping $score : 2^{\mathcal{Q}} \longrightarrow \mathbb{R}$ and the condition*

$$R \subseteq S \;\&\; score(R) > score(T).$$

In our context a score could naturally be for example a linear form in the estimated knowledge space: $score(R) = \sum_{q \in Q} w_q \cdot \mathbb{1}_{f(R)}(q)$ with non-negative weights w_q. If we have two paradoxical response patterns R and S, then the set $A = f(R) \backslash f(S)$ is not empty and we can construct a score as $score(T) = \sum_{q \in A} \mathbb{1}_{f(T)}(q)$ and thus have a paradoxical situation for this score since $score(S) = 0 < score(R)$.

Example 1 (Paradoxical result for a non-quasiordinal basic local independence model (BLIM))**.** *Consider the basic local independence model (BLIM). The BLIM is a quadrupel $(\mathcal{Q}, \mathcal{K}, p, r)$, where*

[3] Statistical preference can be seen as a stochastic relation that is an alternative to stochastic dominance.
[4] See Appendix B.

i) $(Q, \mathcal{K})$ *is a knowledge space with finite* Q*,*

ii) p *is a probability function on* $\mathcal{K}$*, meaning that* $p : \mathcal{K} \longrightarrow [0,1]$ *with* $\sum_{K \in \mathcal{K}} p(K) = 1$*,*

iii) r *is a response function for* $(Q, \mathcal{K}, p)$*, meaning that* $r : 2^Q \times \mathcal{K} \longrightarrow [0,1]$ *with* $\sum_{R \in 2^Q} r(R, K) = 1$ *for arbitrary* $K \in \mathcal{K}$*,*

iv) r *satisfies the condition of local independence:*

$$r(R,K) = \prod_{q \in K \setminus R} \beta_q \cdot \prod_{q \in K \cap R} (1 - \beta_q) \cdot \prod_{q \in R \setminus K} \eta_q \cdot \prod_{q \in Q \setminus (R \cup K)} (1 - \eta_q).$$

Here the β_q *are the probabilities of a careless error and the* η_q *are the probabilities of a lucky guess for each item* q*.*

Now, take the underlying knowledge space $(Q, \mathcal{K})$ *as* $Q = \{q_q, \ldots, q_5\}$ *and* $\mathcal{K} = \{\emptyset, K := \{q_1, q_2, q_4\}, L := \{q_1, q_2, q_3, q_5\}, Q\}$*. Note that* $\mathcal{K}$ *is closed under arbitrary unions but not under arbitrary intersections because* $K \cap L = \{q_1, q_2\} \notin \mathcal{K}$*, thus* $(Q, \mathcal{K})$ *is not a quasi-ordinal knowledge space. Furthermore assume that the careless error and the lucky-guess probabilities are known and equal for all* q*. Thus we will denote them with* β *and* η *respectively and assume that* $0 < \eta < \beta < 0.5$ *which implies in particular that* $(1 - \eta) > (1 - \beta) > 0.5$ *and* $\frac{(1-x)}{x} > 1$ *for* $x \in \{\beta, \eta\}$ *as well as* $\beta(1 - \beta) > \eta(1 - \eta)$*. (These inequalities will be used later.) Finally, take for simplicity for* p *the uniform distribution on all knowledge states* $K \in \mathcal{K}$*. For a given observed response pattern* R *the maximum likelihood estimator for the underlying true knowledge space is simply that* k *in* $\mathcal{K}$ *that maximizes the response value* $r(R, K)$*.*

We are now ready to construct a pair of ***paradoxical response patterns****, namely* $R := \{q_1, q_2\}$ *and* $S := \{q_1, q_2, q_3, q_5\}$*. The following calculations will show that the maximum likelihood estimate of response pattern* R *is* L *and the maximum likelihood estimate of* S *is* K*, but* $L \nsubseteq K$

	q_1	q_2	q_3	q_4	q_5
$\mathcal{Q}$	x	x	x	x	x
K	x	x	x		x
L	x	x		x	
$\emptyset$					
R	x	x			
S	x	x	x		x

$$
\begin{aligned}
r(R,L) &= \beta^1\cdot(1-\beta)^2\cdot\eta^0\cdot(1-\eta)^2 & r(S,K) &= \beta^0\cdot(1-\beta)^4\cdot\eta^0\cdot(1-\eta)^1\\
r(R,\mathcal{Q}) &= \beta^3\cdot(1-\beta)^2\cdot\eta^0\cdot(1-\eta)^0 & r(S,\mathcal{Q}) &= \beta^1\cdot(1-\beta)^4\cdot\eta^0\cdot(1-\eta)^0\\
r(R,K) &= \beta^2\cdot(1-\beta)^2\cdot\eta^0\cdot(1-\eta)^1 & r(S,L) &= \beta^1\cdot(1-\beta)^2\cdot\eta^2\cdot(1-\eta)^0\\
r(R,\emptyset) &= \beta^0\cdot(1-\beta)^0\cdot\eta^2\cdot(1-\eta)^3 & r(S,\emptyset) &= \beta^0\cdot(1-\beta)^0\cdot\eta^4\cdot(1-\eta)^1
\end{aligned}
$$

$$
\begin{aligned}
\frac{r(R,L)}{r(R,\mathcal{Q})} &= \frac{(1-\eta)^2}{\beta^2} = \left(\frac{1-\eta}{\eta}\right)^2 > 1 & \frac{r(S,K)}{r(S,\mathcal{Q})} &= \frac{(1-\eta)}{\beta} > 1\\
\frac{r(R,L)}{r(R,K)} &= \frac{(1-\eta)}{\beta} > 1 & \frac{r(S,K)}{r(S,L)} &= \frac{(1-\beta)^2(1-\eta)}{\beta\eta^2} = \frac{(1-\beta)}{\beta}\cdot\frac{(1-\beta)}{\eta}\cdot\frac{(1-\eta)}{\eta} > 1\\
\frac{r(R,L)}{r(R,\emptyset)} &= \frac{(1-\eta)}{\beta} > 1 & \frac{r(S,K)}{r(S,L)} &= \frac{(1-\beta)^2(1-\eta)}{\beta\eta^2} = \frac{(1-\beta)}{\beta}\cdot\frac{(1-\beta)}{\eta}\cdot\frac{(1-\eta)}{\eta} > 1\\
\frac{r(R,L)}{\emptyset} &= \frac{\beta(1-\beta)^2}{\eta^2(1-\eta)} = \frac{\beta(1-\beta)}{\eta(1-\eta)}\cdot\frac{(1-\beta)}{\eta} > 1 & \frac{r(S,K)}{r(S,\emptyset)} &= \frac{(1-\beta)^4}{\eta^4} = \left(\frac{1-\beta}{\eta}\right)^4 > 1
\end{aligned}
$$

The following theorem shows that the paradox cannot occur if we have a fixed quasi-ordinal knowledge space with known probabilities β_q and η_q.

Theorem 1. *Let $(\mathcal{Q},\mathcal{K},p,r)$ be a basic local independence model where the underlying knowledge space is a quasi-ordinal knowledge space and where the careless-error and the lucky-guess probabilities are known and lie in the interval $(0,0.5)$. Then, the conditioned maximum likelihood estimator*[5]

$$f : 2^{\mathcal{Q}} \longrightarrow \mathcal{K} : R \mapsto f(R) := \operatorname{argmax}_{K\in\mathcal{K}} r(R,K)$$

is not susceptible to the paradox, meaning that there are no two response patterns R and S with a unique ML estimate $f(R)$ and $f(S)$ satisfying

$$R \subseteq S \quad \& \quad f(R) \nsubseteq f(S).$$

Proof. Let R and S be two arbitrary response patterns with $R \subseteq S$ and with unique ML-estimates satisfying $f(R) \nsubseteq f(S)$. We will firstly define some associated sets and illustrate the situation by a small cross tab and sketch the basic idea of the proof before actually doing the proof: Define:

[5] conditioned means here that one maximizes not the unconditional joint likelihood $\mathbb{P}(R = r \;\&\; S = K)$ where R is the random response pattern, r is the actually observed response pattern and S is the random true knowledge state, but one only maximizes $\mathbb{P}(R = r \mid S = K)$ which is equivalent to maximizing the joint likelihood under the assumption that all knowledge spaces are equally probable.

$A := f(R) \cap f(S) \in \mathcal{K}$
$B := f(R) \backslash f(S) \neq \emptyset$
$C := f(R) \cup f(S) = D \dot{\cup} B \in \mathcal{K}$, where
$D := f(S) \subseteq C$

C		x	x	x	x
B		x			
A			x	x	
D=f(S)			x	x	x
$f(R)$		x	x	x	
R	x	x	x		
S	x	x	x	x	

.

Because $A \in \mathcal{K}$ we have $r(R, A) < r(R, f(R)) = r(R, A \dot{\cup} B)$ which means that the addition of the set B to the set A increases the likelihood for the observed response pattern R. In the sequel we show that from this it follows that for a pattern S that is a superset of R the addition of B to the set $f(S)$ will necessarily also increase the likelihood. From $B \dot{\cup} f(S) \in \mathcal{K}$ we can conclude that the knowledge state $f(S)$ cannot be the maximum likelihood estimate of the response pattern S and thus the assumption was wrong which shows that actually the paradox cannot occur:

The fact $r(R, A) < r(R, f(R))$ is equivalent to

$$\prod_{q \in A \backslash R} \beta_q \prod_{q \in A \cap R} (1 - \beta_q) \prod_{q \in R \cap \bar{A} \cap \bar{B}} \eta_q \prod_{q \in R \cap \bar{A} \cap B} \eta_q \prod_{q \in \bar{R} \cap \bar{A} \cap \bar{B}} (1 - \eta_q) \prod_{q \in \bar{R} \cap \bar{A} \cap B} (1 - \eta_q) <$$
$$\prod_{q \in A \backslash R} \beta_q \prod_{q \in B \backslash R} \beta_q \prod_{q \in A \cap R} (1 - \beta_q) \prod_{q \in B \cap R} (1 - \beta_q) \prod_{q \in R \cap \bar{A} \cap \bar{B}} \eta_q \prod_{q \in \bar{R} \cap \bar{A} \cap \bar{B}} (1 - \eta_q),$$

which can be reduced to

$$\prod_{q \in R \cap \bar{A} \cap B} \eta_q \prod_{q \in \bar{R} \cap \bar{A} \cap B} (1 - \eta_q) < \prod_{q \in B \backslash R} \beta_q \prod_{q \in B \cap R} (1 - \beta_q). \tag{1}$$

From (1) it follows that

$$r(S, D) < r(S, \underbrace{C}_{= D \dot{\cup} B}) \tag{2}$$

because (2) is equivalent to

$$\prod_{q\in D\setminus S}\beta_q\prod_{q\in D\cap S}(1-\beta_q)\prod_{q\in S\cap\bar{D}\cap\bar{B}}\eta_q\prod_{q\in S\cap\bar{D}\cap B}\eta_q\prod_{q\in\bar{S}\cap\bar{D}\cap\bar{B}}(1-\eta_q)\prod_{q\in\bar{S}\cap\bar{D}\cap B}(1-\eta_q)<$$
$$\prod_{q\in D\setminus S}\beta_q\prod_{q\in B\setminus S}\beta_q\prod_{q\in D\cap S}(1-\beta_q)\prod_{q\in B\cap S}(1-\beta_q)\prod_{q\in S\cap\bar{D}\cap\bar{B}}\eta_q\prod_{q\in\bar{S}\cap\bar{D}\cap\bar{B}}\eta_q,$$

which can be reduced to

$$\prod_{q\in S\cap\bar{D}\cap B}\eta_q\prod_{q\in\bar{S}\cap\bar{D}\cap B}(1-\eta_q)<\prod_{q\in B\setminus S}\beta_q\prod_{q\in B\cap S}(1-\beta_q). \tag{3}$$

To see that (3) follows from (1) first note that $\bar{D}\cap B=\bar{A}\cap B=B$ and remember that S is a superset of R. Then we can derive (3) from (1) as

$$\begin{aligned}\prod_{q\in S\cap\bar{D}\cap B}\eta_q\prod_{q\in\bar{S}\cap\bar{D}\cap B}(1-\eta_q)&<\prod_{q\in R\cap\bar{A}\cap B}\eta_q\prod_{q\in\bar{R}\cap\bar{A}\cap B}(1-\eta_q)\\&<\prod_{q\in B\setminus R}\beta_q\prod_{q\in B\cap R}(1-\beta_q)\\&<\prod_{q\in B\setminus S}\beta_q\prod_{q\in B\cap S}(1-\beta_q),\end{aligned}$$

where the first and the third inequality can be recognized by observing that we have a product of terms greater than 0.5 (the terms $(1-\beta_q)$ and $(1-\eta_q)$) and terms less than 0.5 (the terms β_q and η_q) and in the product of the left hand sides of the inequalities we have always a subset of terms greater than 0.5 and a superset of terms less than 0.5 compared to the corresponding right hand sights. The second inequality is the inequality from (1).

□

Remark 2. *Note that the theorem assumes that the careless error and lucky guess parameters are known. If they are estimated jointly for all response patterns under the assumption that they are identical for every person (which is actually assumed in the BLIM), then the paradox still cannot occur. If, on the other hand, one estimates the lucky guess and careless error probabilities for each person separately, then the paradox can occur. This is shown in the following example:*

Example 2. *Assume the BLIM with the modification that the careless error probabilities and the lucky guess probabilities are independent from the items but that they are estimated separately for every person via (conditioned) maximum likelihood. The conditioned likelihood for a given response pattern R, an underlying true knowledge state K and careless error- and lucky guess probabilities β and η is given as*

$$\beta^{|K\setminus R|}\cdot(1-\beta)^{|K\cap R|}\cdot\eta^{|R\setminus R|}\cdot(1-\eta)^{|\overline{R\cup K}|}.$$

Assuming $\beta,\eta\in[0,5]$, the likelihood is maximal if

$$\begin{aligned}\beta_{ML}(R,K)&=\begin{cases}\min\{0.5,\frac{|K\setminus R|}{|K\setminus R|+|K\cap R|}\} & \text{if } |K\setminus R|+|K\cap R|\neq 0\\ \in[0,0.5] & \text{if } |K\setminus R|+|K\cap R|=0\end{cases}\\ \eta_{ML}(R,K)&=\begin{cases}\min\{0.5,\frac{|R\setminus K|}{|R\setminus K|+|\overline{K\cup R}|}\} & \text{if } |R\setminus K|+|\overline{K\cup R}|\neq 0\\ \in[0,0.5] & \text{if } |R\setminus K|+|\overline{K\cup R}|=0\end{cases}.\end{aligned}\tag{4}$$

For the knowledge space $(\mathcal{Q},\mathcal{K})$ with $\mathcal{Q}=\{q_1,\ldots,q_5\}$ and $\mathcal{K}=\{\emptyset,K:=\{q_3,q_5\},L:=\{q_1,q_2,q_4\},\mathcal{Q}\}$ and the two response patterns $R=\{q_1,q_4\}$ and $S=\{q_1,q_3,q_4,q_5\}$ we can compute the likelihood of observing such a response pattern (given a certain underlying true knowledge space) under the most likely careless error and lucky guess probabilities from (4):

	q_1	q_2	q_3	q_4	q_5
$\mathcal{Q}$	x	x	x	x	x
K			x		x
L	x	x		x	
$\emptyset$					
R	x			x	
S	x		x	x	x

$$r(R,L) = \beta^1 \cdot (1-\beta)^2 \cdot \eta^0 \cdot (1-\eta)^2 \qquad r(S,K) = \beta^0 \cdot (1-\beta)^2 \cdot \eta^2 \cdot (1-\eta)^1$$
$$r(R,\mathcal{Q}) = \beta^3 \cdot (1-\beta)^2 \cdot \eta^0 \cdot (1-\eta)^0 \qquad r(S,\mathcal{Q}) = \beta^1 \cdot (1-\beta)^4 \cdot \eta^0 \cdot (1-\eta)^0$$
$$r(R,K) = \beta^2 \cdot (1-\beta)^0 \cdot \eta^2 \cdot (1-\eta)^1 \qquad r(S,L) = \beta^1 \cdot (1-\beta)^2 \cdot \eta^2 \cdot (1-\eta)^0$$
$$r(R,\emptyset) = \beta^0 \cdot (1-\beta)^0 \cdot \eta^2 \cdot (1-\eta)^3 \qquad r(S,\emptyset) = \beta^0 \cdot (1-\beta)^0 \cdot \eta^4 \cdot (1-\eta)^1$$

$$\beta_{ML}(R,L) = \frac{1}{3} \quad \eta_{ML}(R,L) = 0 \qquad \beta_{ML}(S,K) = 0 \quad \eta_{ML}(S,K) = 0.5$$
$$\beta_{ML}(R,\mathcal{Q}) = 0.5 \quad \eta_{ML}(S,\mathcal{Q}) = [0,0.5] \qquad \beta_{ML}(S,\mathcal{Q}) = \frac{1}{5} \quad \eta_{ML}(S,\mathcal{Q}) = [0,0.5]$$
$$\beta_{ML}(R,K) = 0.5 \quad \eta_{ML}(R,K) = 0.5 \qquad \beta_{ML}(S,L) = \frac{1}{3} \quad \eta_{ML}(S,L) = 0.5$$
$$\beta_{ML}(R,\emptyset) = (0,0.5) \quad \eta_{ML}(R,\emptyset) = \frac{2}{5} \qquad \beta_{ML}(S,\emptyset) = [0,0.5] \quad \eta_{ML}(S,\emptyset) = 0.5$$

$$r(R,L,\beta_{ML},\eta_{ML}) = \frac{4}{27} = 0.\overline{148} \qquad r(S,K,\beta_{ML},\eta_{ML}) = \frac{1}{8} = 0.125$$
$$r(R,\mathcal{Q},\beta_{ML},\eta_{ML}) = \frac{1}{32} = 0.03125 \qquad r(S,\mathcal{Q},\beta_{ML},\eta_{ML}) = \frac{256}{3125} \approx 0.08192$$
$$r(R,K,\beta_{ML},\eta_{ML}) = \frac{1}{32} = 0.03125 \qquad r(S,L,\beta_{ML},\eta_{ML}) = \frac{1}{27} \approx 0.03703704$$
$$r(R,\emptyset,\beta_{ML},\eta_{ML}) = \frac{108}{3125} \approx 0.03456 \qquad r(S,\emptyset,\beta_{ML},\eta_{ML}) = \frac{1}{32} = 0.03125$$

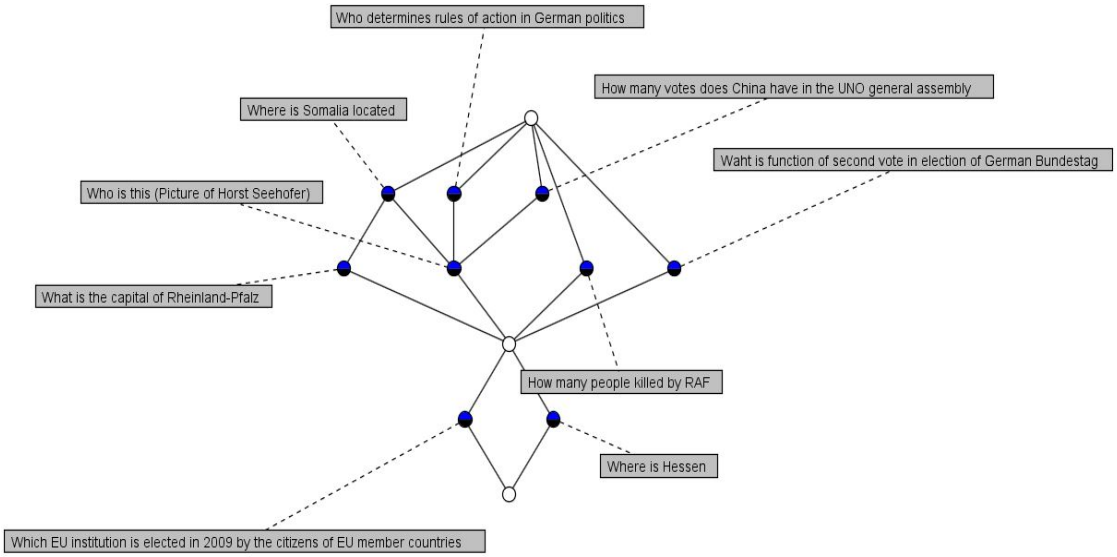
These calculations show that the ML-estimate of pattern R is the knowledge space L and the ML-estimate of the pattern $S \supseteq R$ is $K \not\supseteq L$ which shows that in this situation the paradox is present.

5 Short illustration of the method

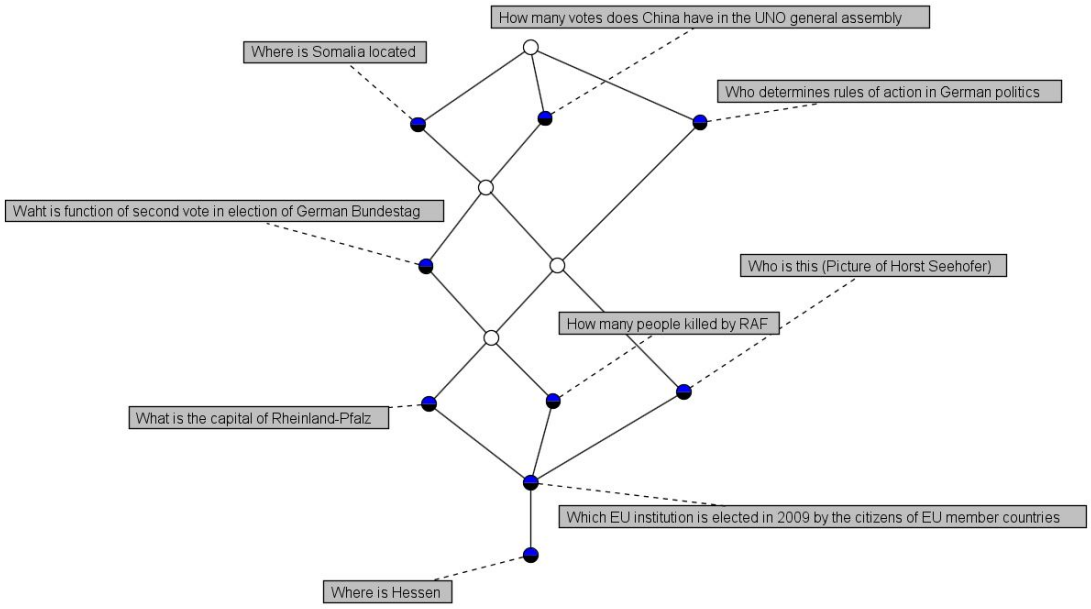
In this section, we shortly show the results of our method for a data set. The used data set is a subsample from the general knowledge quiz *Studentenpisa* conducted online by the German weekly news magazine SPIEGEL ([SPIEGEL Online, 2009]). The data contain the answers of 1075 university students from Bavaria to 45 multiple choice items concerning the 5 different topics *politics*, *history*, *economy*, *culture* and *natural sciences*. For every topic, 9 questions were posed. We compare our method with the minimized corrected inductive item tree analysis algorithm described in Sargin and Ünlü [2009]. For the minimized corrected inductive item tree analysis we used the R package `DAKS` ([Ünlü and Sargin, 2010]). For the computation of the empirically mutually supportive pairs, we used the techniques for detecting stochastic dominance developed in Schollmeyer et al. [2017]. We show here separately for every topic the Hasse graphs of the easiness relation E of the items both for the corrected inductive item tree analysis algorithm as well as for our method. An item i is here more easy than and item j if item i is depicted below item j and if item i is directly or indirectly connected to item j though an ascending path of edges.

topic: Politics

item analysis based on knowledge space theory (IITA algorithm of R package DAKS):

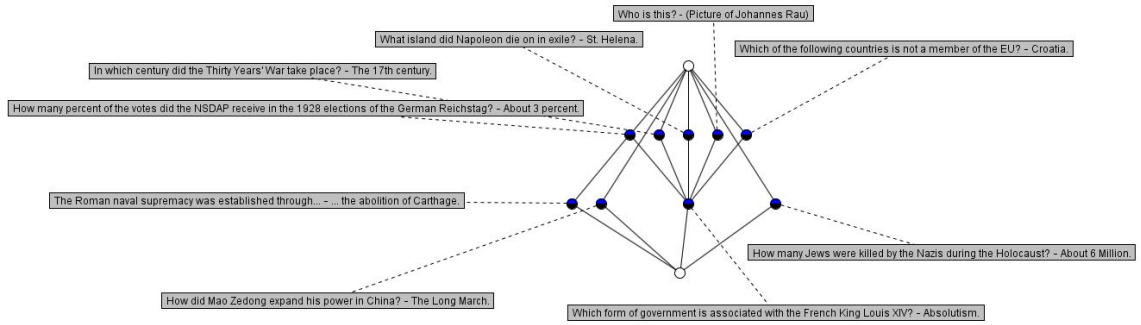


strongest empirically mutually supportive pair (item easiness relation E):

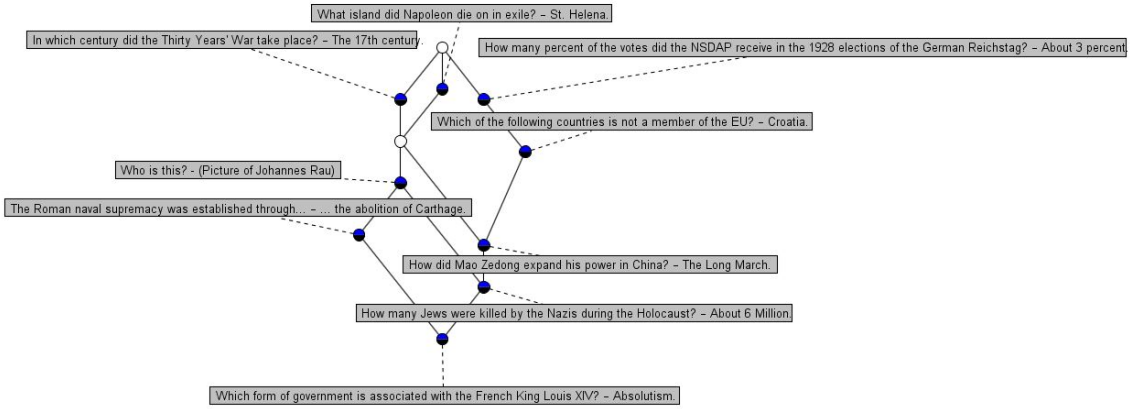


topic: History

item analysis based on knowledge space theory (IITA algorithm of R package DAKS):

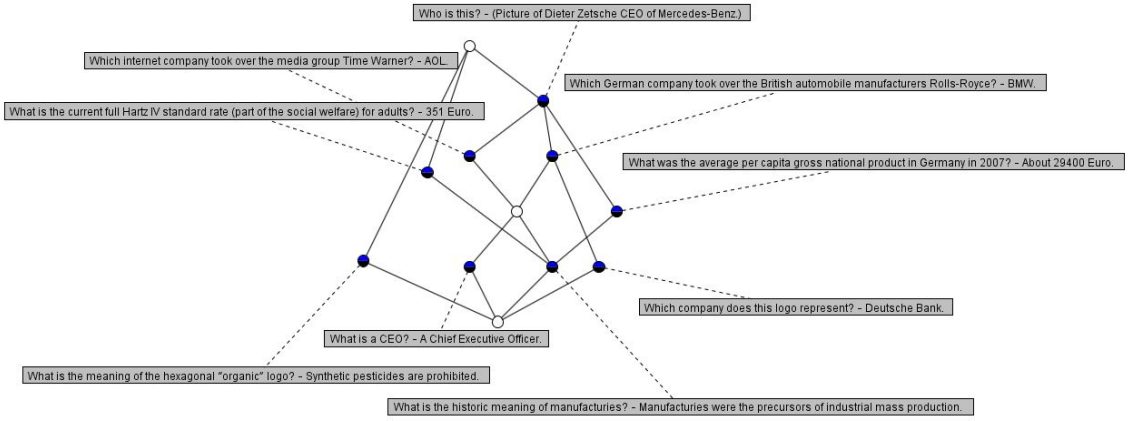


strongest empirically mutually supportive pair (item easiness relation E):

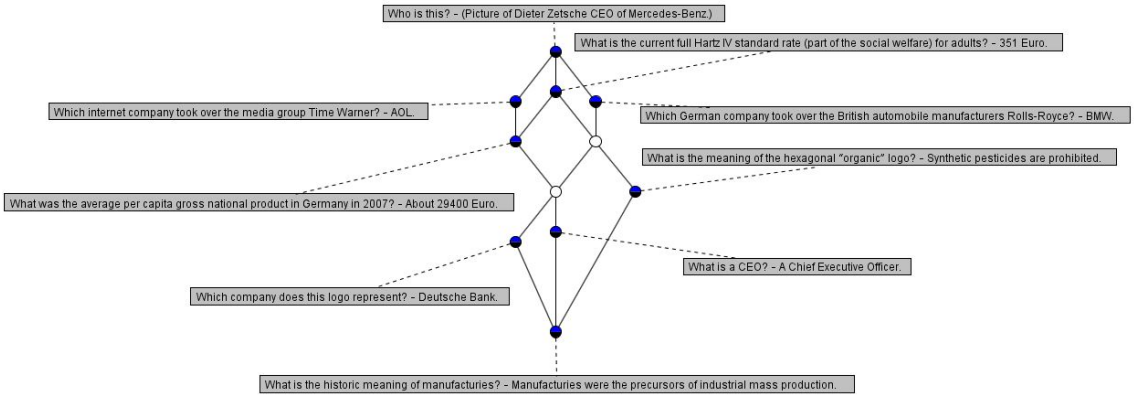


topic: Economy

item analysis based on knowledge space theory (IITA algorithm of R package DAKS):

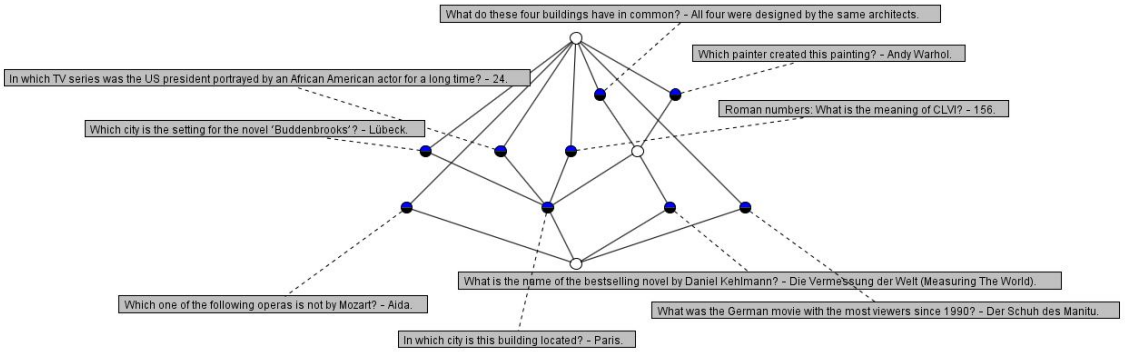


strongest empirically mutually supportive pair (item easiness relation E):

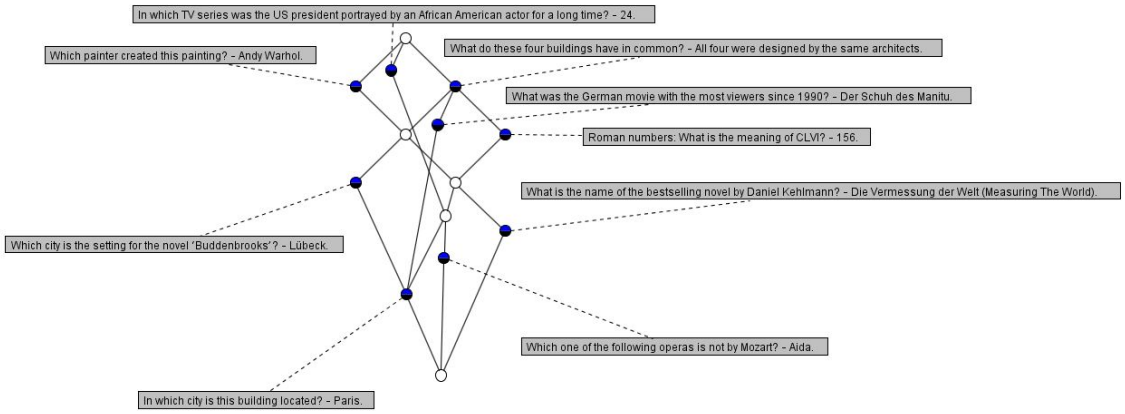


topic: Culture

item analysis based on knowledge space theory (IITA algorithm of R package DAKS):

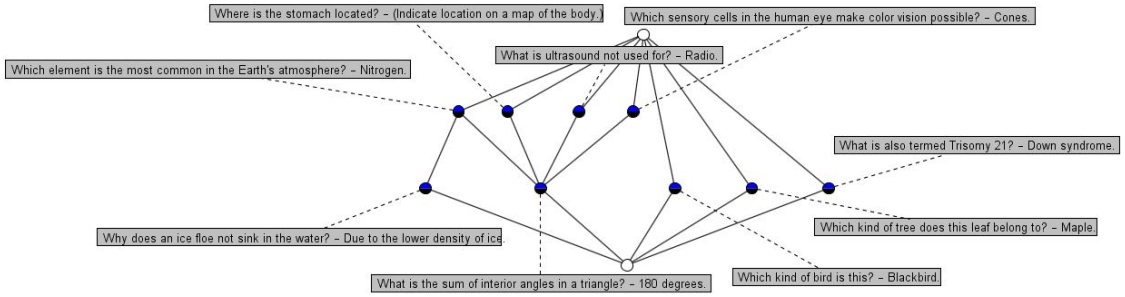


strongest empirically mutually supportive pair (item easiness relation E):

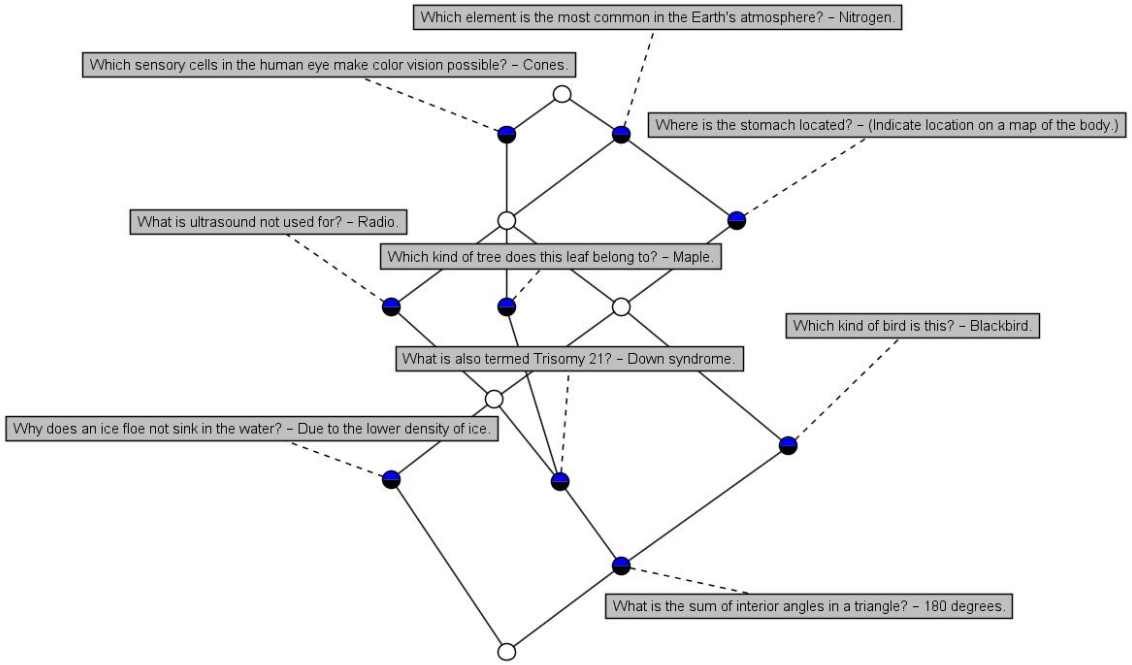


topic: Natural sciences

item analysis based on knowledge space theory (IITA algorithm of R package DAKS):



strongest empirically mutually supportive pair (item easiness relation E):



For our concrete data example, the item easiness relation E of the strongest empirically mutually supportive pair seems to be tendentially stronger than the relation obtained from the inductive item tree analysis. If this is only a coincidence or if this has some reason seems to be not so clear. A further natural question one could ask is how our method statistically behaves under some presumed item response model. Also this question seems to be very difficult to answer in the general multidimensional situation. However, for the case that one has a finite and fixed number of items that are uniformly strictly totally ordered w.r.t. difficulty (meaning that for two different items there is always one item which has smaller solving probabilities, no matter, which person tries to solve it) one can show that if we sample from a population and let the number of sampled persons tend to infinity, then the probability that the observed easiness relation differs from the true difficulty relation goes to zero. The reason for this consistency property is the following:

For two items m_i and m_j where m_i is easier than m_j, the strongest empirically mutually supportive pair will declare m_i as easier than m_j if one can find an appropriate matching of persons. To see that if only n is large enough, with arbitrary high probability one will find such a matching, divide the set of all observed response patterns in all different classes where the responses to all items except the responses to the item m_i and m_j are identical. If n is large enough, then with high probability, in every such class the ordinal relations among the observed frequencies of the different patterns are identical to the ordinal relations among the true probabilities. This means in particular, that with high probability, in every class one observes more persons that solved m_i and not m_j than persons, who solved m_j but not m_i. This means that we can find a matching of persons who solved answers as following: In every class, match persons, who solved both items to itself and match persons, who solved m_j but not m_i to persons who solved m_i but not m_j, which is possible, because there are more persons, who solved m_i but not m_j than persons who solved m_j but not m_i. To see that this matching is showing that m_i is easier than m_j due to the strongest empirically mutually supportive pair, we have to make sure that we have matched persons only to persons who are less successful, but this is clear, because the matched persons did solve the same items up to item m_i and m_j and item m_i was easier than item m_j due to the easiness relation given in the first step of the application of the operator $\mathfrak{L}$.

6 Conclusion

In this paper we have developed a purely descriptive and relational notion of item difficulty and person success. This notion avoids the paradox described in Hooker et al. [2009]. We also shortly indicated, how the descriptive method statistically behaves under certain univariate presumed models of item response theory. For multidimensional models, the behavior of the method seems to be far from clear. Furthermore, also the behavior of the method in comparison to descriptive methods like item tree analysis has still to be further studied.

References

Bolt, D. (2007). The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, 44(4):377–383.

de la Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1):115–130.

De Schuymer, B., De Meyer, H., and De Baets, B. (2003a). A fuzzy approach to stochastic dominance of random variables. In Bilgiç, T., Baets, B. D., and Kaynak, O., editors, *Tenth International Fuzzy Systems Association World Congress*, pages 253–260. Springer.

De Schuymer, B., De Meyer, H., De Baets, B., and Jenei, S. (2003b). On the cycle-transitivity of the dice model. *Theory and Decision*, 54(3):261–285.

DiBello, L. V. and Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, 44(4):285–291.

Doignon, J. and Falmagne, J. (2012). *Knowledge Spaces*. Springer.

Doignon, J.-P. and Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23(2):175 – 196.

Falmagne, J. and Doignon, J. (2010). *Learning Spaces: Interdisciplinary Applied Mathematics*. Springer.

Falmagne, J.-C., Koppen, M., Villano, M., and Doignon, J.-P. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, 97(2):201–224.

Finkelman, M. D., Hooker, G., and Wang, Z. (2010). Prevalence and magnitude of paradoxical results in multidimensional item response theory. *Journal of Educational and Behavioral Statistics*, 35(6):744–761.

Heller, J., Stefanutti, L., Anselmi, P., and Robusto, E. (2015). On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika*, 80(4):995–1019.

Hooker, G., Finkelman, M., and Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika*, 74(3):419–442.

Humphry, S. M. (2013). A middle path between abandoning measurement and measurement theory. *Theory & Psychology*, 23(6):770–785.

Jordan, P. (2013). *Paradoxien in quantitativen Modellen der Individualdiagnostik*. PhD thesis, Universität Hamburg. URL `http://ediss.sub.uni-hamburg.de/volltexte/2013/6198/`.

Jordan, P. and Spiess, M. (2012). Generalizations of paradoxical results in multidimensional item response theory. *Psychometrika*, 77(1):127–152.

Junker, B. W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272.

Kamae, T., Krengel, U., and O'Brien, G. L. (1977). Stochastic inequalities on partially ordered spaces. *The Annals of Probability*, 5(6):899–912.

Lehmann, E. (1955). Ordered families of distributions. *Ann Math Stat*, 26:399–419.

Levhari, D., Paroush, J., and Peleg, B. (1975). Efficiency analysis for multivariate distributions. *The Review of Economic Studies*, 42(1):87–91.

Michell, J. (2008a). Conjoint measurement and the Rasch paradox. *Theory & Psychology*, 18(1):119–124.

Michell, J. (2008b). Is psychometrics pathological science? *Measurement*, 6(1-2):7–24.

Ünlü, A. and Sargin, A. (2010). Daks: An r package for data analysis methods in knowledge space theory. *Journal of Statistical Software, Articles*, 37(2):1–31.

Popper, K. (2005). *The Logic of Scientific Discovery*. Taylor & Francis.

Rasch, G. (1977). On specific objectivity: An attempt of formalizing the generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14:58–94.

Rusch, A. and Wille, R. (1996). Knowledge spaces and formal concept analysis. In Bock, H.-H. and Polasek, W., editors, *Data Analysis and Information Systems: Statistical and Conceptual Approaches Proceedings of the 19th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Basel*, pages 427–436. Springer.

Sargin, A. and Ünlü, A. (2009). Inductive item tree analysis: Corrections, improvements, and comparisons. *Mathematical Social Sciences*, 58(3):376 – 392.

Schollmeyer, G., Jansen, C., and Augustin, T. (2017). Detecting stochastic dominace for poset-valued random variables as an example of linear programming on closure systems. Technical Report 209, Department of Statistics, LMU Munich.

Schrepp, M. (1999). Extracting knowledge structures from observed data. *British Journal of Mathematical and Statistical Psychology*, 52(2):213–224.

Schrepp, M. (2002). Explorative analysis of empirical data by boolean analysis of questionnaires. *Zeitschrift für Psychologie*, 210(2):99–109.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory & Psychology*, 22(6):786–809.

SPIEGEL Online (2009). Studentenpisa - Alle fragen, alle Antworten. In German. accessed 18.08.2017.

Strassen, V. (1965). The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3):337–350.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In Frederiksen, N., Glaser, A., Lesgold, A., and Safto, M., editors, *Diagnostic Monitoring of Skill and Knowledge Acquisition*, pages 453–488. Taylor & Francis.

Ünlü, A. and Sargin, A. (2010). DAKS: An R package for data analysis methods in knowledge space theory. *Journal of Statistical Software*, 37(2):1–31.

Appendix

A Basics of formal concept analysis (FCA)

In formal concept analysis one has given a so-called **formal context** $\mathbb{K} = (G, M, I)$ where G is a set of objects, M is a set of attributes and $I \subseteq G \times M$ is a binary relation between the objects and the attributes with the interpretation $(g, m) \in I$ iff object g has attribute m. If $(g, m) \in I$ we also use infix notation and write gIm. A **formal concept** of the context $\mathbb{K}$ is a pair (A, B) of a set $A \subseteq G$ of objects, called **extent**, and a set $B \subseteq M$ of attributes, called **intent**, with the following properties:

1. Every object $g \in A$ has every attribute $m \in B$ (i.e.: $\forall g \in A \forall m \in B : gIm$).
2. There is no further object $g \in G \backslash A$ that has also all attributes of B (i.e.: $\forall g \in G : (\forall m \in B : gIm) \Longrightarrow g \in A$).
3. There is no further attribute $m \in M \backslash A$ that is also shared by all objects $g \in A$ (i.e. $\forall m \in M : (\forall g \in A : gIm) \Longrightarrow m \in B$).

Conceptually, the concept extent describes, which objects belong to the formal concept and the intent describes, which attributes characterize the concept. The property of being a formal concept can be characterized with the following operators

$$\Phi : 2^M \longrightarrow 2^G : \quad B \mapsto \{g \in G \mid \forall m \in B : gIm\}$$
$$\Psi : 2^G \longrightarrow 2^M : \quad A \mapsto \{m \in M \mid \forall g \in A : gIm\}$$

as

$$(A, B) \text{ is a formal concept } \iff \Psi(A) = B \ \& \ \Phi(B) = A.$$

This can be verbalized as: "The pair (A, B) is a formal concept iff B is exactly the set of all common attributes of the objects of A and A is exactly the set of all objects having all attributes of B."

On the set of all formal concepts one can define a sub-concept relation as

$$(A, B) \leq (C, D) \iff A \subseteq C \ \& \ B \supseteq D.$$

(Actually, for formal concepts the equivalence $A \subseteq C \iff B \supseteq D$ holds.) If the concept (A, B) is a sub-concept of (C, D) then it is a more specific concept containing less objects that have more attributes in common. The set of all formal concepts of a context $\mathbb{K}$ together with the sub-concept relation is called the **concept lattice**. (The concept lattice is in fact a complete lattice.)

A **formal (attribute) implication** is a pair (Y, Z) of subsets of M, also denoted by $Y \longrightarrow Z$. We say that an implication $Y \longrightarrow Z$ is **valid** in a context $\{mathbbK$ if every intent of $\mathbb{K}$ that contains all elements of Y also contains all elements of Z. In this case we also say that the context $\mathbb{K}$ **respects** the implication $Y \longrightarrow Z$. A formal implication $Y \longrightarrow Z$is called **simple** if Y is a singleton.

B Basics of knowledge space theory (KST)

A **knowledge structure** is a tuple $(\mathcal{Q}, \mathcal{K})$ where $\mathcal{Q}$ is a set of questions (items) and $\mathcal{K}$ is a family of subsets of $\mathcal{Q}$ that includes the empty set and the set $\mathcal{Q}$. A set $S \in \mathcal{K}$ is called a knowledge state and models the items a person is able to master. Since mastering an item i may imply mastering an item j, some sets $T \subseteq \mathcal{Q}$ cannot occur as a knowledge state. The set $\mathcal{K}$ models the set of all possible knowledge states, a person could be in. A knowledge structure $(\mathcal{Q}, \mathcal{K})$ where $\mathcal{K}$ is closed under arbitrary unions is called a **knowledge space**. A knowledge space where $\mathcal{K}$ is furthermore closed under arbitrary intersections is called a **quasi ordinal knowledge space**.

C Connections between FCA and KST

The connections between formal concept analysis and knowledge space theory is described in Rusch and Wille [1996]. For a given knowledge space $(\mathcal{Q}, \mathcal{K})$ one can associate the formal context $\mathbb{K} := (G, \mathcal{Q}, I)$, where G is a set of person that answered the set $\mathcal{Q}$ of questions and gIq means that person g did not solve question q. With this association of a formal context to a knowledge space we have a one to one correspondence between formal contexts and knowledge spaces. While knowledge spaces are closed under unions, the concept intents of the associated formal context are the complements of the knowledge states and are closed under intersection. A formal implication $q \longrightarrow r$ could be interpreted as "every person that did not solve item q also did not solve answer r, which could also be stated as mastering item r implies mastering item q.

D Basics of stochastic dominance

For two random variables X, Y on the same probability space $(\Omega, \mathcal{F}, P)$ and with values in a partially ordered set $(V, \leq)$ one says that X is weakly stochastically dominated by Y (w.r.t. first order stochastic dominance) if there exist two copies X' and Y' on another probability space $(\Omega', \mathcal{F}', P')$ with $X \stackrel{d}{=} X'$, $Y \stackrel{d}{=} Y'$ and $P'(X' \leq Y') = 1$. Stochastic dominance can be characterized by the following, essentially equivalent conditions: The random variable X is (weakly) stochastically smaller than the random variables Y if one of the three following conditions is satisfied[6]:

i) $P(X \in A) \leq P(Y \in A)$ for every (measurable) upset $A \subseteq V$

ii) $\mathbb{E}(u \circ X) \leq \mathbb{E}(u \circ Y)$ for every bounded non-decreasing borel-measurable[7] function $u : V \longrightarrow \mathbb{R}$

iii) It is possible to obtain the density[8] f_Y from the density f_X by transporting probability mass from values v to greater or equal values $v' \geq v$.

[6] The equivalence between (ii) and (i) was shown by Lehmann [1955] and independently proved by Levhari et al. [1975]. The equivalence between (iii) and (i) is a consequence of Strassen's Theorem ([Strassen, 1965]), see Kamae et al. [1977].

[7] Here, we have to assume that $(V, \leq)$ can be equipped with an appropriate topology that makes it a partially ordered polish space.

[8] This statement is of course only equivalent if the densities f_X and f_Y actually exist.

Characterization i) is more or less used in definition 1 of this paper, with the difference that one does not have a normalized probability measure P, but instead, a counting measure is underlying. Characterization iii) is very close to the idea of defining a person X as less successful than person Y if one can match every item solved by person X to an item solved by person Y that is at least as difficult.