Diskussionsbeiträge der Projektgruppe Friedensforschung Konstanz, Nr. 74, 2012
www.regener-online.de
ISSN 1611-1818

Wilhelm Kempf

A pragmatic approach to Rasch-modeling: The loss of information index[1]

Abstract: Although attitude questionnaires only rarely satisfy the Rasch Model, the sum score is routinely used in measuring attitudes. This can lead to a considerable loss of diagnostically usable statistical information. As well in Latent Class Analysis, in the Mixed Rasch Model and/or in the Hybrid Model the use of the BIC which favours a smaller number of classes than the AIC, leads to a certain loss of information.
Starting from a general test-theoretical model that includes not only Classical Test Theory but also Item Response Models, the present paper introduces indices for evaluating the relevance of the respective information loss.
An application to the measurement of anti-Semitic attitudes shows that the rejection of the Rasch Model in favor of Latent Class Analysis does imply turning away from the concept of a quantitative attitude dimension. On the other side, however, this provides more accurate information on the structure and dynamics of the respective attitude. Precisely what Classical Test Theory neglects as measuring errors can in the given case contain crucial diagnostic information.

1. Introduction

Attitudes are not directly observable. But they are, however, manifest in expressions of agreement or disagreement with specific statements that can serve as indicators for the respective attitude. We turn this to account in measuring attitudes by presenting a series of statements (items) to subjects, who are asked to indicate agreement or disagreement on a multi-step Likert scale. From the subjects' responses we can then infer their underlying (latent) attitudes, whereby the sum score calculated from the item scores is ordinarily used as an index for the attitude being measured.

At first glance, this procedure seems quite plausible. But it is by no means self-evident that it is really expedient to infer from item responses a (quantitative) latent attitude dimension on which we can order subjects in accord with their sum scores. Thus Kracauer (1952) already pointed out that in content analysis it is not the frequency of specific text characteristics that constitutes the meaning of a text, but rather the patterns into which they are combined, and this may possibly also be the case with attitude measurement.

First of all, although the sum score may provide a (more or less suitable) quantification of the results of a survey, it describes only the subject's manifest responses, but not (yet) the underlying attitude for which questionnaire responses are used as indicators. Secondly, it is not warranted that the sum score represents a usable description of survey results. It is quite conceivable that some responses should be given more weight than others. It is also conceivable that there is no such thing as a general attitude dimension on which subjects can be arranged. Perhaps instead there are various patterns of *qualitatively* different attitudes.

We can regard item responses as indicators of one and the same quantitative attitude dimension if and *only* if each of the items defines the same ordering relation between any two subjects, v and w. Since we must account for the random variation of subjects' responses, it is expedient, however, not to define this ordering relation ($v >^{o} w$) on the basis of the item responses themselves (x_{vi}) but rather, in accordance with $v >^{o} w \Leftrightarrow \tau_{vi} > \tau_{wi}$, on the basis of their expected values $\tau_{vi} = E(X_{vi})$. That the items ($i = 1,\ldots,k$) measure the same attitude dimension can then be inferred from the empirical regularity with which the expected item scores always result in the same rank order among survey subjects independently of the item selection. When this is the case, the questionnaire is *ordinally homogeneous* and the profile lines in Figure 1 do not intersect.[2] If to the contrary the profile lines intersect, the questionnaire is *non-homogeneous*, and it is not possible to attribute the item responses to a common latent dimension.

---

1. English translation of the book chapter "Scorebildung, klassische Testtheorie und Item-Response-Modelle in der Einstellungsmessung" in Wilhelm Kempf & Rolf Langeheine (Hrsg.) (2012). Item-Response-Modelle in der sozialwissenschaftlichen Forschung. Berlin: regener.
Funding provided by the German Research Society (Deutsche Forschungsgemeinschaft – DFG), Grant Number KE 300/8-1.
2. In order to construct profile lines, we transfer the expected item scores of the subjects (τ_{vi}) to the ordinate of a coordinate system on whose abscissa the items are ordered according to the expected total number of points they scored (τ_{oi}). Thereby (from left to right) the item with the highest expected total number of points comes first and that with the lowest comes last.

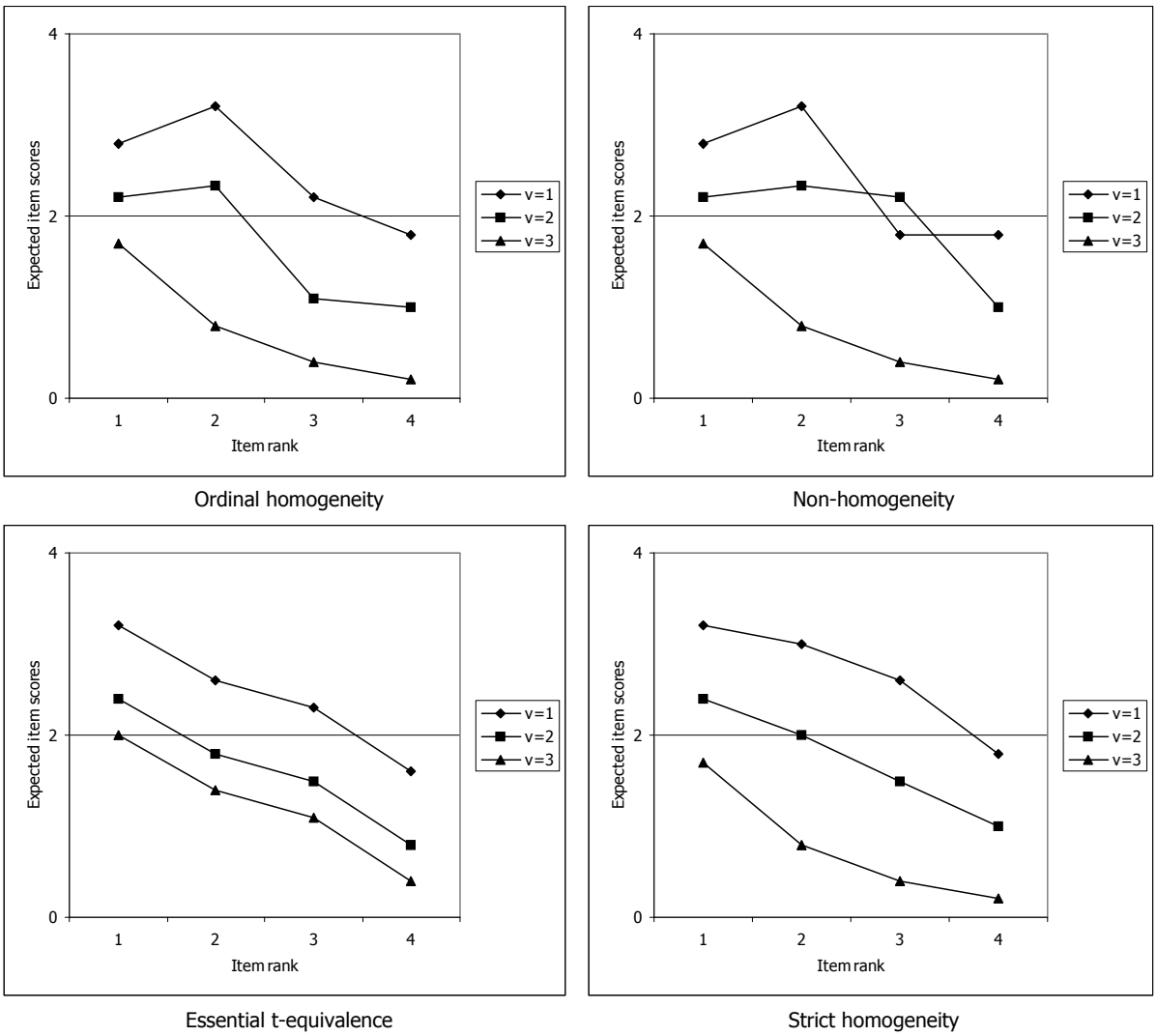

Figure 1: Profile lines of three subjects

Even if a questionnaire is ordinally homogeneous, this still does not mean, however, that the sum scores correctly represent the rank order of the subjects. A necessary (but not sufficient) condition for this is that the items are also ordered according to difficulty, and the rule $i <^{O} j \Leftrightarrow \tau_{vi} > \tau_{vj}$ always defines the same order of difficulty of two items i and j independently of the selection of subjects. Graphically represented, this means that the profile lines shown in Figure 1 are not only non-intersecting, but also monotonically falling.

Classical Test Theory, which we usually draw on in constructing attitude questionnaires, fulfills this precondition by operationalizing the homogeneity of tests or questionnaires through the concept of essential τ-equivalence. This means that the expected scores of any two items differ merely through a subject-independent constant $c_{ij} = \tau_{vi} - \tau_{vj}$, which expresses the difference in the difficulty of the two items (cf. Lord & Novick, 1968: 50, 135). If this is the case, the profile lines shown in Figure 1 are parallel.

Since the profile lines for any given pair of subjects must be parallel, this must also hold for each pair of – however defined –classes of subjects. This introduces the possibility of statistically testing the essential *τ-equivalence* of the items by means of simple t-tests and/or ANOVA. In everyday questionnaire construction, however, we usually forgo such model tests and instead use a more pragmatic procedure. Since we can hardly assume that the items are essentially τ-equivalent, we try instead to quantify how well (or poorly) the essential τ-equivalence of the items – and thereby the *internal consistency* of a questionnaire – is assured.

In fact, Cronbach's coefficient alpha, which we use for this, is upwardly bounded by the reliability of the questionnaire. The more the essential τ-equivalence is violated, the lower it becomes, and the more it is fulfilled the closer it comes to reliability. In order to serve as a reliable measure for the internal consistency of a questionnaire it would

thus have to be related to the reliability of the questionnaire. But since still more rigorous preconditions apply to all other measures of reliability, and since they can both over- and underestimate reliability when these preconditions are not fulfilled, this is not possible (cf. Kempf, 2008: 173f).[1] Apart from the coefficient alpha itself, which represents a lower bound for reliability, Classical Test Theory therefore cannot give a trustworthy measure of reliability.

But it is not just this deficiency in Classical Test Theory that makes the usual procedures of questionnaire construction appear methodologically unsatisfactory. As has been known since Rasch's precedent-setting work (1960, 1961), score construction can involve a considerable loss of statistical information. The sum score fully exploits the potential information in the subjects' response patterns about the latent person variable (= the underlying attitude) if (and only if) the probability density of the response variables can be represented by the logistic function of the Rasch Model. Only in this case does the sum score represent a sufficient statistic for the characteristics of the latent person variable. The profile lines are then, to be sure, likewise non-intersecting and falling monotonically, but they are not parallel (cf. Figure 1), for which reason internal consistency is from the start not a suitable criterion to justify score construction.

Especially in attitude measurement the sufficiency of the sum score is, however, only rarely fulfilled. Attitude questionnaires are only rarely *strictly homogeneous* in the sense of the Rasch Model. Nevertheless, we routinely rely on score construction and are satisfied if a questionnaire displays adequate internal consistency in the sense of Cronbach's coefficient alpha.

Even if there are pragmatic arguments in favor of this procedure, it nevertheless raises the question of how large the entailed loss of statistical information is. Or, put differently: If we want to employ score construction even though a questionnaire is not strictly homogeneous, shouldn't we at least specify a measure of relevance that quantifies the extent of the information loss connected with it? And – if the information loss proves too great – shouldn't we forgo score construction and conceive of the latent person variable as not a quantitative, but rather a qualitative variable, as is the case with Latent Class Analysis – going back to Lazarsfeld (1950) –, by means of which we can identify different *types* of response patterns.

2. A general test-theoretical model

In order to construct such a measure of relevance, we start from a general test-theoretical model (Kempf, 2008: 199ff) that includes not only Classical Test Theory, but also the Item Response Models. Although classical and stochastic test theories arose in different research traditions, they are based on compatible assumptions that can be expressed, analogous to Novick's (1966) basic assumptions of Classical Test Theory, in the following model assumptions:

1. Each subject v and each item i corresponds to a discrete random variable X_{vi} with the range x = 0,..., m-1 and the probability density $f_{vi}(x)$.
2. The items are locally independent, so that the probability of the response vector of a given subject can be represented by the formula:

$$\text{prob}(x_{v1},...,x_{vk}) = \prod_{i=1}^{k} f_{vi}(x_{vi}) . \quad (1)$$

If the responses of n subjects are available, they can be represented in the form of a response matrix $X = ((x_{vi}))$. The lines of the response matrix form the so-called response vectors $x_v = (x_{v1},\ldots,x_{vk})$, which reproduce the response pattern of a subject and thereby provide a complete description of his test results. This response pattern represents the data basis from which we can infer the characteristics of the latent person variable underlying the responses.

In *Classical Test Theory* the form of the density function $f_{vi}(x)$ is not further specified. Instead, merely the expected value of the response variables $\tau_{vi} = E(X_{vi})$, designated as the true score, is modeled according to

$$x_{vi} = \tau_{vi} + f_{vi} , \quad (2)$$

whereby $f_{vi} = x_{vi} - \tau_{vi}$ describes the measuring error with which the subject's response is burdened.

In *Stochastic Test Theory*, in contrast, additional assumptions are made about the probability density of the response variables, and depending on what assumptions are made thereby, various Item Response Models can be differentiated, among which there is a conceptual hierarchy, represented in Figure 2.

1. The only exception to this is retest reliability, which always underestimates the reliability of a test when the preconditions for its calculation are not fulfilled.

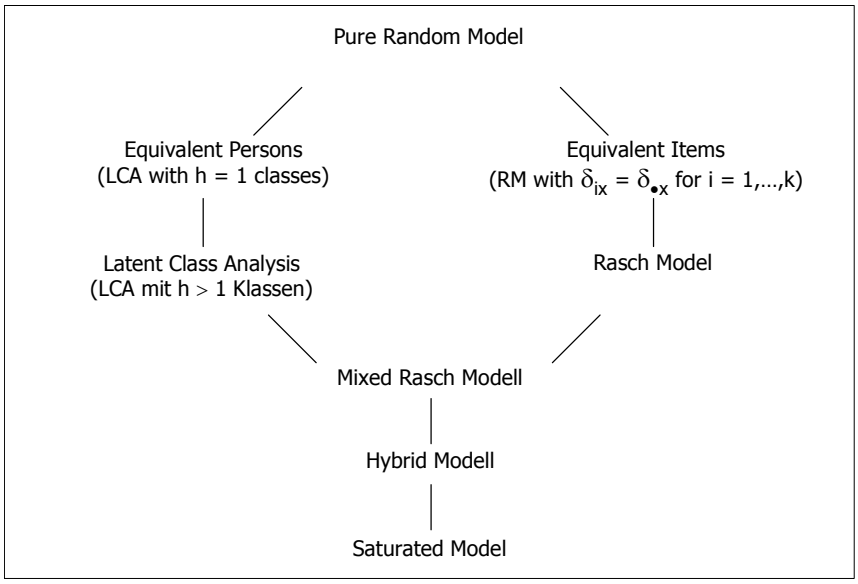

Figure 2: The conceptual hierarchy of the Item Response Models (according to Kempf, 2008: 200; modified)

Pure random variation: The pure random model (PR) postulates that the response variables X_{vi} are distributed independently not only of persons, but also of items, so that all subjects and items have the same response variables with the probability density

$$f_{vi}(x) = f_{\bullet\bullet}(x) = p_{\bullet\bullet x} \,. \tag{3}$$

If this assumption is fulfilled, the questionnaire provides no information about either differences in attitudes between subjects or differences in difficulty between items. In this case, it is unusable as a diagnostic instrument. All variations in the data are purely random variations, and the (marginal) likelihood of the response matrix has the form

$$L_{PR} = \prod_{x=0}^{m-1} p_{\bullet\bullet x}^{n_{oox}} \quad \text{with} \quad n_{oox} = \sum_{v=1}^{n}\sum_{i=1}^{k} n_{vix} \,, \tag{4}$$

whereby

$$n_{vix} = \begin{cases} 1 \text{ if } x_{vi} = x \\ 0 \text{ otherwise} \end{cases} . \tag{5}$$

Although the PR model (still) does not trace subjects' responses back to a latent person variable and assumes no differences between items, it is nevertheless of theoretical interest insofar as the likelihood of the PR model represents a lower bound of the likelihood that can be obtained with Item Response Models.

Equivalent persons: Likewise, the questionnaire does not contain any diagnostic information on differences in attitudes between subjects when the response variables of different items are dissimilar but are distributed independently of the person, so that all subjects have the same response variables, with the probability density

$$f_{vi}(x) = f_{\bullet i}(x) = p_{\bullet ix} \,. \tag{6}$$

Formally seen, this model resembles the one-class solution of Latent Class Analysis (LCA).

Latent person variables: The questionnaire contains diagnostic information about the subjects only if the response variables are distributed depending on the person. If this is the case, the probability density of the response variables can be represented as a function of the latent person variable Θ

$$f_{vi}(x) = f_i(x, \theta_v) \,, \tag{7}$$

which can be either one- or multi-dimensional. For the moment, only the uni-dimensional case is of interest, whereby we can again distinguish between discrete and continuous person variables.

Latent trait models: If the latent person variable is continuous, it is measured on a metric scale with the range $-\infty \leq \theta_v \leq \infty$. The probability that a randomly selected subject v will answer item i in the category x can then be represented in the form

$$\text{prob}(X_{vi} = x) = \int_{-\infty}^{\infty} f_i(x,\theta)\, g(\theta)\, d\theta\,, \tag{8}$$

whereby $g(\theta)$ designates the probability density of the latent variable.

Equivalent items: In the simplest case, there are no differences between the various items, so that all items possess the same response variables with the probability density

$$f_i(x,\theta_v) = f_\bullet(x,\theta_v) = p_{v\bullet x}\,. \tag{9}$$

Formally viewed, this model corresponds to the Rasch Model (RM), with $\delta_{ix} = \delta_{\bullet x}$ for i = 1,…, k and m = 0,…, m-1.

The Rasch Model: Basically we could use any continuous distribution function for $f_i(x,\theta_V)$. Only the logistic function

$$f_{vi}(x,\theta_v) = \frac{\exp(x\theta_v - \delta_{ix})}{\sum_{y=0}^{m-1}\exp(y\theta_v - \delta_{iy})} \quad \text{with } \delta_{i0} = 0, \tag{10}$$

as assumed in the RM, however, assures that the sum score will completely exploit the statistical information on the latent person variable which is contained in the response vector. Therein, θ_v designates the value of the latent person variable, and δ_i the difficulty of answering item i in category x.

If the RM is fulfilled, the sum score groups the subjects into h = k*(m-1) +1 (manifest) classes – measured on an ordinal scale – x_{vo} = 0,…,k*(m-1), so that the marginal likelihood of the response matrix

$$L_{RM} = \prod_{v=1}^{n} \text{prob}\{(x_{v1},\ldots,x_{vk})|x_{vo}\} \prod_{g=0}^{h} p_g^{n_g} \tag{11}$$

can be divided into two factors, the first of which – so-called conditional likelihood – is independent of the latent person variable. Therein n_g designates the number of subjects whose score equals g, and p_g designates the probability that a randomly chosen subject will have this score.

Latent Class Analysis: If the latent person variable is discrete, it is measured on a nominal or ordinal scale with the range $\theta_g = \theta_1,\ldots,\theta_h$, which divides the subject pool into h (latent) classes g = 1,…,h, so that $\theta_v = \theta_g \Leftrightarrow v \in g$, and all subjects who belong to the same class possess the same class-specific category probabilities

$$f_i(x,\theta_g) = p_{gix}\,. \tag{12}$$

The marginal likelihood of the response matrix then takes the form

$$L_{LCA} = \prod_{v=1}^{n}\left(\sum_{g=1}^{h} p_g \prod_{i=1}^{k}\prod_{x=0}^{m-1} p_{gix}^{n_{vix}}\right), \tag{13}$$

whereby the class size p_g again represents the probability that a randomly chosen subject belongs to class g.

The Mixed Rasch Model represents a combination of quantitative and qualitative latent person variables that goes back to Rost (1990), Mislevy & Verhelst (1990) and Keldermann & Macready (1990) and divides the subjects into h latent classes to which the Rasch Model applies.

Performance measurement is a typical area of application of the MRM, and to be sure when test problems can be solved using several different solution strategies. Depending on the strategy that subjects use to solve the problem, different abilities are called for, and problems that are relatively easy to solve with one strategy can prove difficult using another strategy and vice versa. The MRM then allows a two-step diagnostic process: In the first step we can diagnose the latent class the subjects belong to (i.e., what solution strategy they use), and in the second step the members of the same class can be compared in terms of their performance level (i.e., in regard to how well they can use this strategy). In attitude measurement the results of the MRM, however, can often be interpreted only with difficulty, if at all.

The same holds for the so-called *Hybrid Models* (HM) (e.g. Davier & Rost, 1997), which likewise divide subjects into h latent classes to which the RM applies and beyond this also permit a residual class of subjects whose response patterns are incompatible with the RM. In terms of our example, these could be subjects who do not use a uniform solution strategy, but rather switch back and forth between various strategies.

Saturated model: If each possible response pattern defines its own class of subjects with $\theta_g \Leftrightarrow (x_{g1},...,x_{gk})$, then

$$f_i(x,\theta_g) = \begin{cases} 1 & \text{for } x = x_{gi} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

Since the subjects' response patterns can be directly observed, the person variable Θ is in this case no longer latent, but rather manifest, and the marginal likelihood of the response matrix

$$L_{sat} = \prod_{g=1}^{h_{max}} p_g^{n_g}, \quad \text{with} \qquad h_{max} = m^k, \tag{15}$$

depends only on the class size parameters p_g. Since the saturated model does not contain any restrictive model assumptions, it provides (in absolute terms) the best possible description of the data, and the marginal likelihood in equation 15 cannot be surpassed by any of the other models. In view of m^k possible response patterns, the saturated model is, however, also interpretable only in exceptional cases. It is, nonetheless, of theoretical interest insofar as the likelihood of the saturated model represents the upper bound of the likelihood that can be achieved by Item Response Models.

3. Information-theoretical measures of relevance

A crucial difference between the RM and LCA is that LCA does not fulfill any homogeneity preconditions, and the decision for the characterization of the latent person variables is not based on the sum score alone, but rather on subjects' complete response patterns. As well, no preliminary decision is made about whether the classes differ quantitatively or qualitatively. For these reasons, LCA is also suitable for the description of data if the RM applies: With the ideal model validity of the RM and a suitable number of classes of LCA, the latent classes identified by LCA are identical with the manifest classes which the RM constructs on the basis of the subjects' sum scores. Much the same holds – although on a more complex level – for the MRM and the HM.

The hierarchy of models represented in Figure 2 is only conceptual, therefore. Viewed formally, all other models represent a special case of LCA, whereby the latent classes coincide in part with the manifest classes. The higher a model is located on the formal hierarchy of models (cf. Figure 3), the more restrictive are its model assumptions.

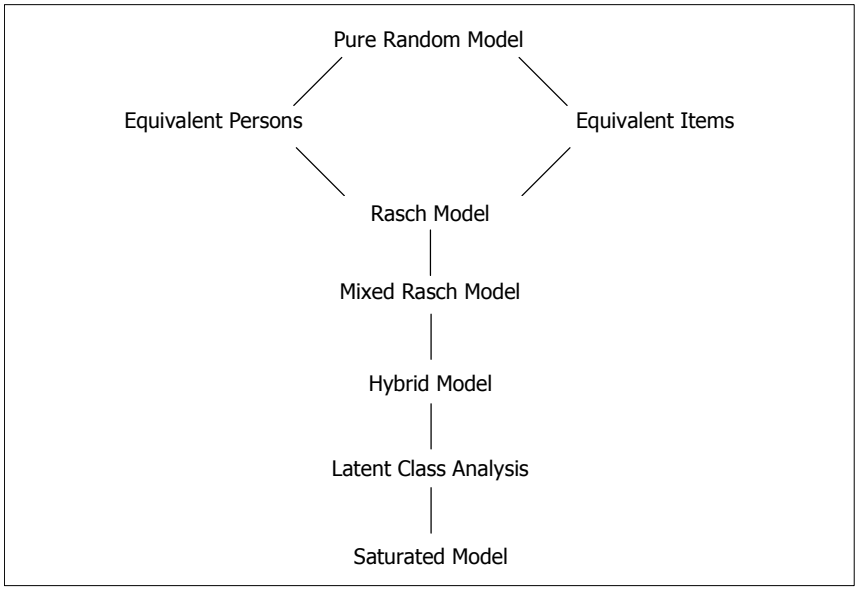

Figure 3: The formal hierarchy of the Item Response Models

As a measure of relevance for the loss of statistical information that goes together with score construction, it is therefore expedient to relate the explanatory power of the RM to that of LCA.

3.1 Explanatory-Power Index

In order to construct such a measure, we proceed in several steps and first calculate the (absolute) redundancy of the data (measured in bit pro response pattern) which is identified by a given model (MX):

$$C_{abs,MX} = H_{max} - H_{MX} . \tag{16}$$

Therein H_{MX} refers to the entropy (= the mean information) of a response pattern if the MX model applies, which in accord with

$$H_{MX} = -\frac{1}{n \ln(2)} \ln(L_{MX}) \tag{17}$$

can be represented as a function of the likelihood (L_{MX}) that the response matrix has if the MX model applies (cf. Appendix 1).

H_{max} refers to the maximum possible entropy of a response pattern, which can be calculated according to

$$H_{max} = k \frac{\ln(m)}{\ln(2)} \tag{18}$$

(cf. Appendix 2) and assumes that all response categories possess the same person- and item-independent *a-priori* probability $p_{\bullet\bullet x} = m^{-1}$, so that all possible response patterns m^k are equally probable.

Finally, inserting (17) and (18) into (16) yields, for the redundancy of the data identified by the model, the formula

$$C_{abs,MX} = \frac{n\ k \ln(m) + \ln(L_{MX})}{n \ln(2)} \tag{19}$$

(cf. Appendix 3). The greater the value this expression assumes, the more information the model exploits.

Since already the assumption of empirical category probabilities ($p_{\bullet\bullet x}$) that can deviate from the a-priori probability (m^{-1}) exploits a certain amount of statistical information, it is reasonable to calculate in a second step the difference

$$VI_{MX} = C_{abs,MX} - C_{abs,PR} \tag{20}$$
$$= \frac{\ln(L_{MX}) - \ln(L_{PR})}{n \ln(2)} ,$$

which measures the model-specific (= going beyond the assumption of empirical category probabilities) statistical information that the respective model (MX = RM or LCA) exploits. Because $L_{PR} \leq L_{MX} \leq L_{sat}$, the expression in equation 20 can, however, assume at most the value

$$VI_{max} = C_{abs,sat} - C_{abs,PR} \tag{21}$$
$$= \frac{\ln(L_{sat}) - \ln(L_{PR})}{n \ln(2)} ,$$

which gives the statistical information that can be *maximally* exploited by assuming a latent person variable.

By relating the expressions in equations (20) and (21), we thereby obtain a measure of the (relative) explanatory power of the MX model which specifies the share of the exploitable information that is exploited by the respective model.

$$EP_{MX} = \frac{VI_{MX}}{VI_{max}} \tag{22}$$
$$= \frac{\ln(L_{MX}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})} .$$

3.2 Loss-of-Information Index

The so-defined EP index is constructed analogous to the well-known PRE measure (proportional reduction in error) in sociological statistics, which was introduced by Goodman (1972) and draws on work by Goodman & Kruskal (1954). Reynolds (1977) points out that PRE does not give any information about whether the respective model adequately explains the data, but rather merely compares the fit of two models (here of the PR and the MX model) with each other. We can easily point to cases in which the PRE or respectively EP are in fact high, but the likelihood ratio test against the saturated model is nevertheless significant. Precisely this, however, qualifies the measure as a basis for the quantification of the information loss which goes together with score construction if the RM doesn't fit. For this purpose we recommend the Loss of Information index

$$LI_{RM,LCA} = 1 - \frac{EP_{RM}}{EP_{LCA}} \qquad (23)$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

(cf. Appendix 4), which describes the share of statistical information that continues to be unexploited with the assumption of strict homogeneity (RM), in contrast to LCA (no homogeneity assumption):

Data example 1: As an example, let us take a questionnaire for the measurement of manifest anti-Semitism including the following three items

1. "One shouldn't do any trade and commerce with Jews."
2. "It is better to have nothing to do with Jews."
3. "I am one of the people who do not like any Jews."

This was answered by N = 411 subjects on a 5-step Likert scale which ranges from complete disagreement (x = 0) to complete agreement (x = 4). Since one of the subjects answered the items incompletely, the following calculations are based on a sample of n = 410.

The internal consistency of the scale amounts to α = 0.836. By means of LCA, four latent classes of subjects were identified using Akaike's (1987) information criterion (AIC) (cf. Table 1). The likelihood ratio test of the 4-class model as opposed to the saturated model speaks for a very good model fit (L-ratio = 33.98; df = 73; n.s.), and the explanatory power of the model is more than satisfactory (EP = 94.31%).

Model	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR	4	-1249,82	597,64	120	< 0,001	2507,64	0,00%
LC1	12	-1221,16	540,32	112	< 0,001	2466,32	9,59%
LC2	25	-1026,76	151,52	99	< 0,001	2103,52	74,65%
LC3	38	-986,76	71,52	86	n.s.	2049,52	88,03%
LC4	51	-967,99	33,98	73	n.s.	2037,98	94,31%
LC5	64	-964,30	26,60	60	n.s.	2056,60	95,55%
LC6	77	-958,99	15,98	47	n.s.	2071,98	97,33%
RM	23	-1028,36	154,72	101	< 0,001	2090,72	74,11%
Saturated	124	-951,00				2150,00	100,00%

Table 1: Data example 1, Goodness-of-fit statistics

The profile lines of the latent classes are non-intersecting and (cum crano salis) fall monotonically (cf. Figure 4). Nevertheless, the RM must be rejected (L-ratio = 154.72; df = 101; p < 0.001). The scale is *not* strictly homogeneous, but rather only ordinally homogeneous. The explanatory power of the RM is rather unsatisfactory (EP = 74.11%), and the sum score construction is accompanied by a considerable loss of statistical information (LI = 21.42%).

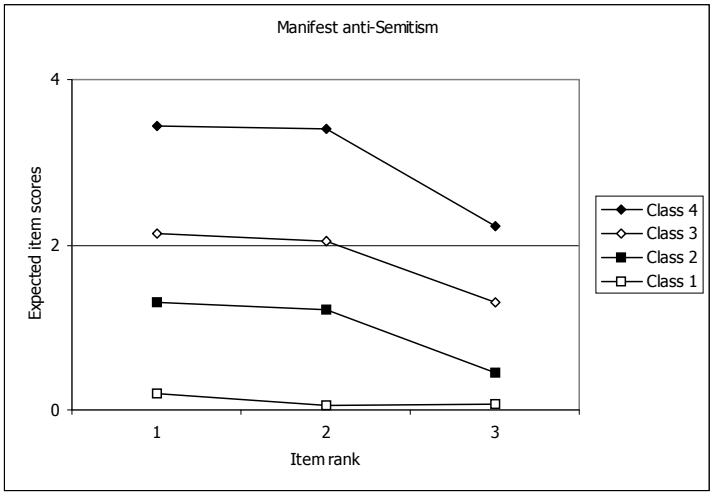

Figure 4: Data example 1, Profile lines of the latent classes

3.3 Corrected Loss-of-Information Index

Often we have to deal with missing data when making questionnaire surveys. To be able to construct sum scores there is no choice other than to either – as in Data example 1 – exclude subjects from the data analysis who answered the questionnaire incompletely or to recode the missing responses in the neutral category ("neither-nor").

Data example 2: As an example of data recoded in this way, let us examine a questionnaire for the measurement of secondary anti-Semitism with the three items:

1. "Decades after the end of the war we should not talk so much about the persecution of the Jews, but rather finally close the books on the past."
2. "One should ultimately put an end to the chitchat about our guilt vis-à-vis the Jews."
3. "The German people (does not have) a particular responsibility vis-à-vis the Jews."

These were answered by n = 411 subjects on the same 5-step Likert scale as the items in Data example 1.

The internal consistency of the scale amounts to $\alpha = 0.772$. To show that the calculations are based on recoded data, in Table 2 the models are marked with an asterisk (*). As the Table shows, LCA, in accord with AIC, identified five latent classes. The likelihood-ratio test of the 5-class model, as opposed to the saturated model, speaks for a good model fit (L-ratio = 52.99; df = 60; n.s.), and the explanatory power of the model is satisfactory (EP = 91.77%).

Model	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR*	4	-1933,93	643,52	120	< 0.001	3875,86	0,00%
LC1*	12	-1899,37	574,41	112	< 0.001	3822,74	10,74%
LC2*	25	-1742,37	260,41	99	< 0.001	3534,74	59,54%
LC3*	38	-1693,22	162,12	86	< 0.001	3462,45	74,81%
LC4*	51	-1668,18	112,03	73	< 0.005	3438,36	82,59%
LC5*	64	-1638,66	52,99	60	n.s.	3405,32	91,77%
LC6*	77	-1636,28	48,22	47	n.s.	3426,55	92,51%
RM*	23	-1728,14	231,95	101	< 0.001	3502,28	63,96%
Saturated*	124	-1612,17				3472,34	100,00%

Table 2: Data example 2, Goodness-of-fit statistics

The RM, in contrast, must be rejected (L-ratio = 231.95; df = 101, p < 0.001), its explanatory power is unsatisfactory (EP = 63.96%), and the sum score construction in this example is accompanied by a dramatic loss of diagnostically relevant statistical information (LI = 30.30%).

As the profile lines of the latent classes represented in Figure 5 show, the scale is to be sure ordinally homogeneous, so that each of the three items define the same rank order between each of two classes (non-intersecting profile

lines). However, the profile lines do not run monotonically, which indicates that the difficulty relation between the items shifts with increasingly strong anti-Semitic attitudes. While subjects who (rather) disapproved of the secondary anti-Semitic statements (Classes 1 and 2) clearly disapproved more strongly of the claim that Germans should finally stop the chitchat about guilt vis-à-vis the Jews (Item 2) than of the simple demand to close the books that is expressed in 1, it is just the opposite with hardened anti-Semites (Class 5). They agree more strongly with the statement in Item 2, which not only calls for closing the books on the past, but also rejects the guilt question as mere "chitchat" and thereby borders on Holocaust denial. This information, which is highly relevant for the diagnosis of anti-Semitic attitudes, is lost, however, if we merely consider the subjects' sum scores.

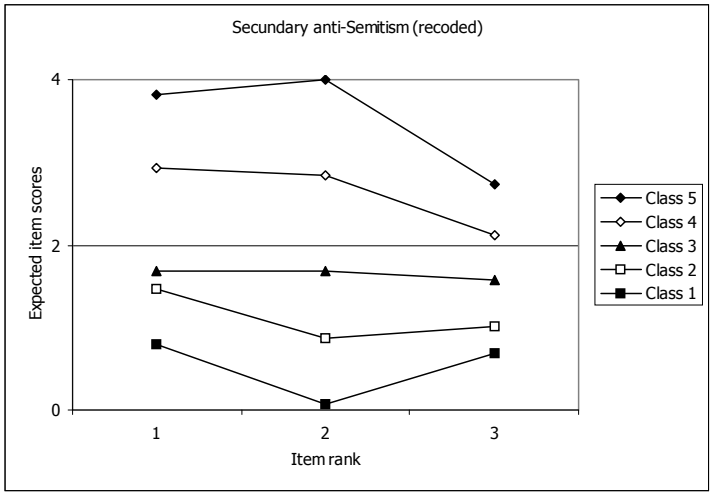

Figure 5: Data example 2, Profile lines of the latent classes

Information is lost not only through the exclusion of incompletely filled out questionnaires (cf. Data example 1), but also through the recoding of missing data (cf. Data example 2). To be sure, subjects who leave an item unanswered *neither* agree *nor* disagree with the statement, but rather than placing the item in the 'neither-nor' category they simply do not check any answer.

Since this response behavior can very well have its own diagnostic meaning, it is reasonable to do recoding only for Rasch analysis and to calculate the LCA using missing data as an additional response category (x = m).

Because the number of response categories (= m) underlying Rasch analysis is smaller in this case, however, than of those underlying LCA (= m+1), not only the likelihood of the pure random model, but also that of the saturated model has different values in the reduced data format $L^*_{PR} \neq L_{PR}$ or respectively $L^*_{sat} \neq L_{sat}$ than in the data format underlying LCA. In order to indicate that the likelihood and the explanatory power of the RM are also calculated with the reduced response format, we likewise mark them with an asterisk and write L^*_{RM} or respectively EP^*_{RM}. Likewise we mark the maximally exploitable information through the assumption of a latent person variable in the reduced data format with VI^*_{max}.

In order to be able to compare the two models (and data formats), we can then use a corrected loss-of-information index

$$LI^*_{RM,LCA} = 1 - \frac{EP^*_{RM}}{EP_{LCA}} \frac{VI^*_{max}}{VI_{max}} \qquad (24)$$

$$= \frac{\ln(L_{LCA}) - \ln(L^*_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})} - \frac{\ln(L_{PR}) - \ln(L^*_{PR})}{\ln(L_{LCA}) - L(L_{PR})}$$

(cf. Appendix 5).

If LCA and RM are calculated on the basis of the same response format, then $\ln(L^*_{PR}) = \ln(L_{PR})$, and the correction factor $\frac{\ln(L_{PR}) - \ln(L^*_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})} = 0$ drops out, so that $LI^*_{RM,LCA}$ is reduced to $LI_{RM,LCA}$.

Data example 3: as an example let us look again at the above-mentioned questionnaire for the measurement of secondary anti-Semitism and when doing LCA treat declining to answer as an independent response category (x = 5), but do Rasch analysis, to the contrary, using recoded data, as above.

As the goodness-of-fit statistics in Table 3 show, in this case as well LCA identified five latent classes (in accord with AIC) whose profile lines in the present case scarcely differ from those in Data example 2 (cf. Figure 5). The likelihood ratio test, as opposed to the saturated model, again shows a very good model fit (L-ratio = 59.44; df = 136; n.s.). The explanatory power of LCA is slightly higher than in Data example 2 (EP = 90.91%). Also the loss of statistical information (LI* = 30.74%) that goes with *recoding and score construction* is only slightly greater than it appears in Data example 2.

This is, however, not an argument that speaks against taking missing data into account as an independent response category in LCA and applying the corrected loss-of-information index. The triviality of the differences between Data examples 2 and 3 is plainly attributable to the fact that very little missing data was contained in the available data.

Model	n(P)	ln(L)	L-Ratio	df	p	AIC	EP
PR	5	-1940,41	653,72	210	< 0.001	3890,82	0,00%
LC1	15	-1904,49	581,88	200	< 0.001	3838,98	10,99%
LC2	31	-1746,76	266,42	184	< 0.001	3555,52	59,25%
LC3	47	-1697,21	167,32	168	n.s.	3488,42	74,40%
LC4	63	-1671,45	115,80	152	n.s.	3468,90	82,29%
LC5	79	-1643,27	59,44	136	n.s.	3444,54	90,91%
LC6	95	-1636,77	46,44	120	n.s.	3463,54	92,90%
Saturated	215	-1613,55				3657,10	100,00%
PR*	4	-1933,93	643,52	120	< 0.001	3875,86	0,00%
RM*	23	-1728,14	231,94	101	< 0.001	3502,28	63,96%
Saturated*	124	-1612,17				3472,34	100,00%

Table 3: Data example 3, Goodness-of-fit statistics

In order to represent the profile lines of the latent classes, we can in this case take two approaches. One consists in calculating the expected item scores with recoded data (no response = neither-nor). This is, however, not very satisfactory. It is better to calculate the expected item scores solely on the basis of the data of those subjects who did *not* refrain from answering a respective item. This can be done very simply by projecting the class-specific category probabilities p_{gix} according to

$$p^{\#}_{gix} = \frac{p_{gix}}{\sum_{y=0}^{h} p_{gix}} \quad \text{for } x = 0,\ldots,m\text{-}1 \tag{25}$$

to 100% and recalculating the expected item scores according to

$$\tau_{gi} = \sum_{x=0}^{m-1} x\, p^{\#}_{gix}\,. \tag{26}$$

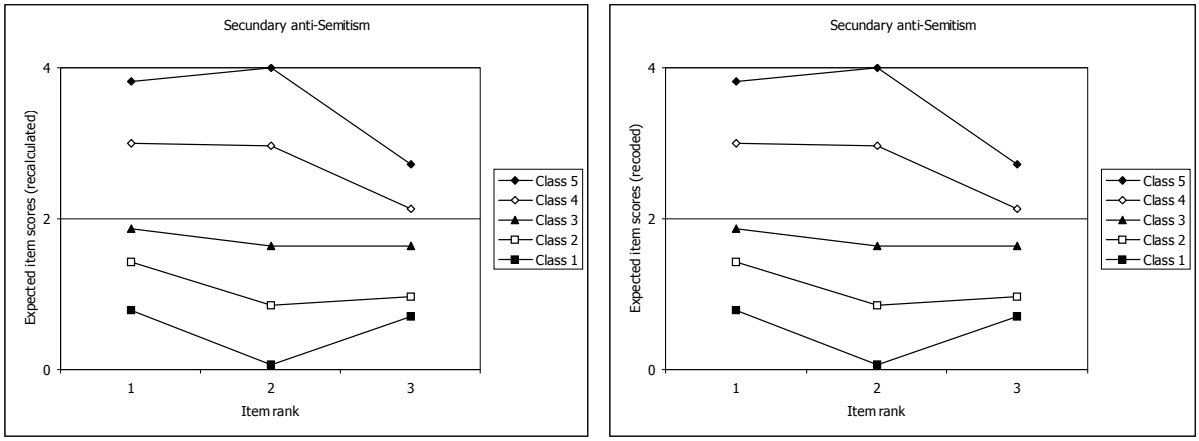

Figure 6: Data example 3, Profile lines of the latent classes

In our Data example, the two types of calculation make hardly any difference (cf. Figure 6), which among other things is again attributable to the fact that the data contained very little missing data.

4. Model selection

In determining the number of latent classes suitable for the description of the data, in LCA, MRM and/or in HM we proceed quite pragmatically. Since the likelihood of the response matrix increases monotonically with an increasing number of classes, we cannot simply select the number of classes for which the response matrix possesses the greatest likelihood. If we did this, it would always lead to a choice in favor of the saturated model.

The more latent classes we assume, the more model parameters we have to estimate from the data. The more model parameters we have to estimate, the less the distribution of the response matrix will be restricted, and the less the distribution of the response matrix is restricted the greater will be its likelihood.

As a criterion for the goodness of the description, we therefore select a measure like the AIC index

$$AIC = 2\left|\ln(L_{MX})\right| + 2\ n(P_{MX})\ , \tag{27}$$

which relates the likelihood of the data matrix to the number of model parameters that need to be estimated. Thereby L_{MX} designates the likelihood, and $n(P_{MX})$ the number of independent model parameters of the respective model.

Since the likelihood is a number between zero and one, the absolute value of its natural logarithm is smaller the greater the likelihood is. The AIC index is therefore smaller the more precisely the data are described (high likelihood) and the more parsimonious the description (low number of parameters).

Therefore as the best model we select the number of classes that minimizes the AIC index and thereby represents an optimal compromise between exactitude and parsimony.

In large samples, however, the AIC index easily inflates the number of classes, for which reason in this case we use other measures of information, such as the BIC index, going back to Bozdogan (1987)

$$BIC = 2\left|\ln(L_{MX})\right| + \ln(n)\ n(P_{MX})\ , \tag{28}$$

which gives the number of classes more weight and thereby favors a smaller number of classes. As an example for this, in Table 4 we contrast the AIC and BIC indices of the above-discussed data examples. As the table shows, in all three cases the BIC does in fact favor a smaller number of classes.

	Data example 1			Data example 2			Data example 3		
Model	AIC	BIC	EP	AIC	BIC	EP	AIC	BIC	EP
LC1	2466,32	2514,51	9,59%	3822,74	3870,96	10,74%	3838,98	3899,26	10,99%
LC2	2103,52	2203,92	74,65%	3534,74	3635,20	59,54%	3555,52	3680,10	59,25%
LC3	2049,52	2202,13	88,03%	3462,45	3615,15	74,81%	3488,42	3677,29	74,40%
LC4	2037,98	2242,80	94,31%	3438,36	3643,31	82,59%	3468,90	3722,07	82,29%
LC5	2056,60	2313,63	95,55%	3405,32	3662,51	91,77%	3444,54	3762,01	90,91%
LC6	2071,98	2381,22	97,33%	3426,55	3735,99	92,51%	3463,54	3845,31	92,90%
$LI_{BIC,AIC}$	6,66%			18,48%			18,15%		

Table 4: Comparison of the AIC and BIC indices

At what point a sample becomes large remains rather vague, however. Are n = 410 (Data example 1) or n = 411 (Data examples 2 and 3) already large samples? With this sample size, is BIC already preferable to AIC? Doesn't the choice of the smaller number of classes favored by BIC (Model MY) lead to a relevant loss of information in contrast to what AIC favors (Model MX)?

To decide which of the two information measures we should rely on in the concrete case it is again helpful to calculate a loss-of-information index that contrasts the explanatory power of the two models:

$$LI_{MY,MX} = 1 - \frac{EP_{MY}}{EP_{MX}} \qquad (29)$$

$$= \frac{\ln(L_{MX}) - \ln(L_{MY})}{\ln(L_{MX}) - \ln(L_{PR})}.$$

Applied to our data examples (cf. Table 4), it appears that the choice of BIC is connected with a very considerable loss of information in two cases, $LI_{AIC,BIC}$ = 18.48% (Data example 2) and respectively $LI_{AIC,BIC}$ = 18.15% (Data example 3). If in these two cases we choose the 3-class solution proposed by BIC, also the information advantage of LCA, as opposed to the RM, of $LI_{RM,AIC}$ = 30.30% decreases to $LI_{RM,BIC}$ = 14.50% (Data example 2) or respectively of $LI^*_{RM,AIC} = 30.74\%$ to $LI^*_{RM,BIC} = 15.38\%$ (Data example 3). In both data examples, it is thus advisable to use the 5-class solution suggested by AIC.

In Data example 1, the information loss that goes with a choice in favor of BIC is, to the contrary, with $LI_{AIC,BIC}$ = 6.66% not really relevant (LI < 10%) and clearly less than the loss resulting from the sum score ($LI_{RM,LCA}$ = 21.42%). If the lower number of classes suggested by BIC can be interpreted better than the number of classes that AIC suggests, in cases like this we can decide in favor of BIC. For the quantification of the information loss that goes with sum score construction, we should, however, in each case start from the number of classes suggested by AIC. Otherwise the information loss would be underestimated (in our example with $LI_{RM,BIC}$ = 16.55% instead of $LI_{RM,AIC}$ = 21.42%).

5. Summary

The usual employment of sum scores in attitude measurement is tied to preconditions that are seldom checked in practice. If these preconditions are not fulfilled, the sum score produces a more or less serious loss of diagnostically relevant statistical information with regard to the latent person variable.

In fact this information is fully exploited by the sum score only if the questionnaire is strictly homogeneous in the sense of the RM. If this precondition is not fulfilled, we should at least be able to give a measure of relevance that quantifies the extent of the information loss resulting from use of the sum score. If the information loss proves to be too great, we should refrain from score construction and not conceive of the latent person variable as a quantitative attitude dimension, but rather as a qualitative variable.

For this purpose, in the present paper an information-theoretical loss-of-information index (equations 23) is introduced that confronts the explanatory power of the RM with that of LCA, which unlike the RM does not meet any homogeneity preconditions.

While LCA is able to model missing data as an independent response category, with the RM this is not the case. In order to be able to construct sum scores, we must then either exclude from the data any subjects who did not fill out the questionnaire completely or recode the missing responses in the neutral category ("neither-nor"). Information is likewise lost not just by excluding incompletely filled-out questionnaires, but also by recoding missing data. In order to quantify the loss of information that goes with recoding and sum score construction, a corrected loss-of-information index (equation 24) was therefore developed that takes account of the missing data in LCA as an independent response category and calculates the RM only on the basis of recoded data.

If in the interpretation of the results of LCA, of the MRM or the HM we do not rely on the number of classes favored by AIC, but rather on the (lower) number of classes suggested by BIC, the loss of information that goes with this can likewise again be measured using a loss-of-information index (equation 29).

Appendix

Appendix 1: The entropy of a response pattern with the adoption of the MX model.

According to the Shannon-Wiener formula, the entropy of a response pattern is calculated from

$$H_{MX} = -\sum_{\tilde{y}} p_{\tilde{y}} \, ld\left(p_{\tilde{y}}\right),$$

whereby the sum sign covers all possible response patterns $\left(\tilde{y}\right)$, and $ld\left(p_{\tilde{y}}\right)$ designates the binary logarithm of the probability $\left(p_{\tilde{y}}\right)$ of the response pattern. Expressed in natural logarithms, this equation can also be written in the form

$$H_{MX} = -\frac{1}{n}\sum_{\tilde{y}} n \, p_{\tilde{y}} \, \frac{\ln\left(p_{\tilde{y}}\right)}{\ln(2)}$$

$$= -\frac{1}{n \, \ln(2)}\sum_{\tilde{y}} n_{\tilde{y}} \, \ln\left(p_{\tilde{y}}\right),$$

wherein $n_{\tilde{y}}$ refers to the frequency of the response pattern and the formula

$$\sum_{\tilde{y}} n_{\tilde{y}} \, \ln\left(p_{\tilde{y}}\right) = \ln\left\{\prod_{\tilde{y}} p_{\tilde{y}}^{n_{\tilde{y}}}\right\}$$

is equal to the natural logarithm of the likelihood of the response matrix, so that

$$H_{MX} = -\frac{1}{n \, \ln(2)} \ln\left(L_{MX}\right).$$

Appendix 2: The maximal entropy of a response pattern

Given k items with m response categories there are m^k possible response patterns, whose maximal entropy

$$H_{max} = ld\left(m^k\right),$$

due to $ld\left(m^k\right) = k \, ld(m)$, and after recalculating into a natural logarithm, can also be expressed in the form

$$H_{max} = k \, \frac{\ln(m)}{\ln(2)} \, .$$

Appendix 3: The redundancy of a response vector with the adoption of the MX model

Inserting for H_{max} and H_{MX} in

$$C_{abs,MX} = H_{max} - H_{MX}$$

we obtain

$$C_{abs,MX} = k \, \frac{\ln(m)}{\ln(2)} + \frac{1}{n \, \ln(2)} \ln\left(L_{MX}\right)$$

$$= \frac{n \, k \, \ln(m) + \ln\left(L_{MX}\right)}{n \, \ln(2)} \, .$$

Appendix 4: The loss-of-information index

Inserting for EP_{RM} and EP_{LCA} in

$$LI_{RM,LCA} = 1 - \frac{EP_{RM}}{EP_{LCA}},$$

we obtain

$$LI_{RM,LCA} = 1 - \frac{\frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}}{\frac{\ln(L_{LCA}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}}$$

$$= 1 - \frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}.$$

Because of

$$\frac{\ln(L_{LCA}) - L(L_{PR})}{\ln(L_{LCA}) - L(L_{PR})} = 1,$$

this expression can also be stated in the form

$$LI_{RM,LCA} = \frac{\ln(L_{LCA}) - L(L_{PR})}{\ln(L_{LCA}) - L(L_{PR})} - \frac{\ln(L_{RM}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{PR}) - \ln(L_{RM}) + \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

$$= \frac{\ln(L_{LCA}) - \ln(L_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})}.$$

Appendix 5: The corrected loss-of-information index

Inserting for EP^*_{RM}, EP_{LCA}, VI^*_{max} and VI_{max} in

$$LI^*_{RM,LCA} = 1 - \frac{EP^*_{RM}}{EP_{LCA}} \frac{VI^*_{max}}{VI_{max}},$$

we obtain

$$LI^*_{RM,LCA} = 1 - \frac{\frac{\ln(L^*_{RM}) - \ln(L^*_{PR})}{\ln(L^*_{sat}) - \ln(L^*_{PR})}}{\frac{\ln(L_{LCA}) - \ln(L_{PR})}{\ln(L_{sat}) - \ln(L_{PR})}} \frac{\frac{\ln(L^*_{sat}) - \ln(L^*_{PR})}{n \ln(2)}}{\frac{\ln(L_{sat}) - \ln(L_{PR})}{n \ln(2)}}$$

$$= 1 - \frac{\ln(L^*_{RM}) - \ln(L^*_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}.$$

Because of

$$\frac{\ln(L_{LCA}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})} = 1$$

we can also write this expresssion in the form

$$LI^{*}_{RM,LCA} = \frac{\ln(L_{LCA}) - \ln(L_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})} - \frac{\ln(L^{*}_{RM}) - \ln(L^{*}_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})}$$

$$= \frac{\ln(L_{LCA}) - \ln(L^{*}_{RM})}{\ln(L_{LCA}) - \ln(L_{PR})} - \frac{\ln(L_{PR}) - \ln(L^{*}_{PR})}{\ln(L_{LCA}) - \ln(L_{PR})} \ .$$

References

Akaike, Hirotugu (1987). Factor Analysis and AIC. Psychometrika, 52, 317-332.

Bozdogan, Hamporsum (1987). Model selection for Akaike's information criterion (AIC). Psychometrika, 53, 345-370.

Davier, Matthias v. & Rost, Jürgen (1997). Self-Monitoring – A class variable?, in: Rost, Jürgen & Langeheine, Rolf (eds.). Applications of latent trait and latent class models in the social sciences, 296-304.

Goodman, Leo A. (1972). A modífied multiple regression approach to the analysis of dichotomous variables. American Sociological Review, 37, 28-46.

Goodman, Leo A. & Kruskal, William H. (1954). Measures of association for cross-classifications. Journal of the American Statistical Association, 54, 310-364.

Kelderman, Henk, & Macready, George B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. Journal of Educational Measurement, 27, 307-327.

Kempf, Wilhelm (2008). Forschungsmethoden der Psychologie. Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik. Band 2: Quantität und Qualität. Berlin: regener.

Kracauer, Siegfried (1952). The challenge of qualitative content analysis. Public Opinion Quaterly, 16, 631-642.

Lazarsfeld, Paul F. (1950). Logical and mathematical foundations of latent structure analysis. In: Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A. & Clausen, J.A. (eds.). Studies in social psychology in world war II, Vol. IV. Measurement and Prediction. Princeton/N.J.: Princeton University Press, 362-412.

Lord, Frederic M. & Novick, Melvin R. (1968). Statistical theories of mental test scores. Reading (Mass.): Addison-Wesley.

Mislevy, Robert J. & Verhelst, Norman (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrica, 55, 195-215.

Novick, Melvin R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

Rasch, Georg (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rasch, Georg (1961). On general laws and the meaning of measurement in psychology. Berkeley: University of California Press.

Reynolds, Henry T. (1977). The analysis of cross-classifications. London:The Free Press.

Rost, Jürgen (1990). Rasch models in latent classes: An integration of two approaches to item analysis. Applied Psychological Measurement, 14, 271-282.