Developments in Social Science Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 18, Ljubljana: FDV, 2002

On the Assessment of Gain Scores by Means of Item Response Theory

Gerhard H. Fischer[1]

Abstract

The problem of the measurement and statistical assessment of change based on test scores which arises in repeated measurement designs with two time points is treated within an item response theory framework. The latter is delineated by postulating a Partial Credit Model, of which the Rating Scale Model and the Rasch Model are special cases. A conditional maximum likelihood estimator of the amount of change, Clopper-Pearson and related significance tests for the change parameter, uniformly most accurate confidence intervals, and uniformly most powerful unbiased tests are presented. They are all 'exact' in the sense that no asymptotic approximations are needed. They are grounded on the conditional distribution of the gain score, given the sum score of both time points. These methods are quite flexible because they do not require the same test to be given on both occasions; it is necessary, though, that the items presented at the two time points be chosen from an item pool conforming to the Partial Credit Model, and that the item $\times$ category parameters of that model be known (i.e. have been estimated with sufficient precision from a previous sample of testees). Possible applications are the computation of significance tables for 'gain scores' (i.e. score differences) for fixed pre and posttests or the computation of the significance of a score difference for an individual in individualized (adaptive) testing. That all results hold for single individuals and thus are applicable in single case studies is a noteworthy feature of the present methods.

1 Motivation

The measurement and evaluation of change in testees between testing occasions is highly important in psychological assessment. Questions concerning the amount of

[1] University of Vienna, Department of Psychology, A 1010 Wien (Vienna), Liebiggasse 5, Austria. e-mail: gh.fischer@univie.ac.at
This work was supported in part by Dr. Schuhfried Ges.m.b.H., Mödling, Austria.

change in testees' abilities or traits invariably occur in many subfields of psychology, especially in clinical, developmental, applied, and social psychology. Nevertheless, unsolved methodological problems of measuring change were notorious already in classical psychometrics (Harris, 1963; Cronbach and Furby, 1970; Willett, 1989): 'change scores' (or 'gain scores') $D = R_2 - R_1$, where R_1 and R_2 are raw scores in a test of interest, observed on two occasions, were considered inherently *unreliable*. The problem of their reliability has recently been discussed again by Williams and Zimmerman (1996). Amongst other results, these authors have shown that

$$\mathrm{Rel}(D) = \frac{\lambda\,\mathrm{Rel}(R_1) + \lambda^{-1}\mathrm{Rel}(R_2) - 2\,\rho(\tau_1,\tau_2)\sqrt{\mathrm{Rel}(R_1)\mathrm{Rel}(R_2)}}{\lambda + \lambda^{-1} - 2\,\rho(\tau_1,\tau_2)\sqrt{\mathrm{Rel}(R_1)\mathrm{Rel}(R_2)}}, \qquad (1.1)$$

where 'Rel' denotes reliability, λ the quotient $\sigma(R_1)/\sigma(R_2)$, and $\rho(\tau_1,\tau_2)$ the correlation of the true scores between the time points (testing occasions) T_1 and T_2. Therefore, $\mathrm{Rel}(D)$ is a function of (i) pretest reliability, (ii) posttest reliability, (iii) the pretest-posttest true score correlation, and (iv) the quotient λ.

As can be seen from Table 1, when R_1 and R_2 have similar standard deviations, i.e, for λ near 1.0, $\mathrm{Rel}(D)$ decreases dramatically as a function of $\rho(\tau_1,\tau_2)$, especially when $\rho(\tau_1,\tau_2)$ is close to 1.0. (By setting the numerator in (1.1) to zero it can be shown that, in the present example, $\mathrm{Rel}(D)$ becomes zero at $\lambda = 1.06$, which value, however, is not given in Table 1.) Cases where λ is in the neighborhood of 1.0 are likely to occur in test practice, and high true score correlations are often considered desirable (as an indication that the same trait is measured at T_1 and T_2); but in precisely these cases, the reliability of the difference score D necessarily becomes low.

This – and similar – considerations have nurtured the suspicion that difference (or gain) scores are poor measures of change. Some authors have even advised avoiding gain scores altogether (Cronbach and Furby, 1970).

These discouraging results are not germane to difference scores *per se*, they are rather caused by the employment of test reliability as an indicator of measurement precision. Indeed, it is quite common wrongly to consider reliability a characterization of test precision, which it is *not*: reliability, defined as the quotient of true score variance over total variance of a measure (cf. Lord and Novick, 1968, p. 61), strongly depends on distributional properties of the reference population, which have nothing to do with measurement precision. Therefore, the fact that – under certain conditions – the reliability of the gain score becomes small does *not* imply that the gain score cannot be used as a basis for the estimation of true change and/or is uninformative with respect to testing the H_0 of no change. As will be shown below, within the framework of a suitably chosen family of IRT models the gain score can well be used (i) to derive a measure of change with quite favorable properties, (ii) to construct a confidence interval for the true amount of change, and (iii) to test the H_0 of no change. Thus, from an IRT perspective, we shall arrive at an assessment

Table 1: Reliability of gain score $D = R_2 - R_1$, for $\mathrm{Rel}(R_1) = .800$ and $\mathrm{Rel}(R_2) = .900$, as a function of λ and $\rho(\tau_1, \tau_2)$.

	$\rho(\tau_1,\tau_2)$										
λ	.000	.100	.200	.300	.400	.500	.600	.700	.800	.900	1.000
.10	.899	.897	.895	.894	.892	.890	.888	.886	.883	.881	.879
.20	.896	.893	.889	.885	.881	.876	.871	.865	.859	.853	.846
.30	.892	.886	.881	.874	.867	.859	.850	.839	.827	.813	.797
.40	.886	.879	.871	.862	.851	.839	.825	.807	.786	.760	.726
.50	.880	.871	.861	.849	.835	.818	.798	.771	.737	.692	.626
.60	.874	.863	.851	.837	.819	.798	.770	.734	.685	.612	.497
.70	.867	.856	.842	.825	.805	.779	.745	.699	.633	.530	.345
.80	.861	.848	.833	.815	.792	.763	.724	.669	.588	.455	.193
.90	.855	.842	.826	.806	.781	.750	.707	.646	.555	.398	.073
1.00	.850	.836	.819	.799	.773	.739	.694	.631	.533	.365	.010
1.20	.841	.827	.809	.788	.761	.727	.681	.618	.521	.361	.038
1.40	.834	.819	.802	.781	.755	.722	.679	.621	.536	.401	.158
1.60	.828	.814	.797	.777	.753	.722	.683	.631	.559	.452	.275
1.80	.824	.810	.794	.775	.752	.724	.689	.644	.584	.498	.369
2.00	.820	.807	.792	.774	.753	.728	.696	.657	.606	.537	.440
3.00	.810	.800	.788	.776	.761	.745	.726	.705	.679	.649	.613
4.00	.806	.798	.789	.779	.769	.757	.745	.731	.715	.697	.677
5.00	.804	.797	.790	.783	.774	.766	.756	.746	.735	.722	.709
6.00	.803	.797	.791	.785	.778	.771	.764	.756	.747	.738	.728
7.00	.802	.797	.792	.787	.781	.775	.769	.763	.756	.748	.740
8.00	.802	.797	.793	.788	.783	.778	.773	.768	.762	.756	.749
9.00	.801	.797	.794	.789	.785	.781	.776	.771	.766	.761	.756
10.00	.801	.798	.794	.790	.787	.783	.779	.774	.770	.766	.761

of the precision of change measurements that is very different from that of Williams and Zimmerman (1976).

2 The partial credit model for a repeated measurement design

The Partial Credit Model (PCM; Masters, 1982) is an IRT model for unidimensional polytomous items with ordered response categories (also denoted as 'graded response items'). It is the most general model within a family of Rasch Models, comprising the Rating Scale Model (RSM; Rasch, 1965; Andrich, 1978) and the Rasch Model (RM; Rasch, 1960) as special cases. Therefore, choosing the PCM as a basis implies

that the results will be applicable both to unidimensional achievement tests with dichotomous items (with response categories 'right' vs. 'wrong') or with polytomous items (e.g., with response categories 'right', 'partially right', and 'wrong') and to unidimensional rating scales with ordered response categories (like 'always', 'often', 'rarely', 'never'). In the latter case, the response categories may either be defined differently for each item (admissible in the PCM) or identically for all items (as required by the RSM). Therefore, the methods presented below have a wide spectrum of applications.

The PCM for a set of items given to a testee S at one time point (say, T_1) is defined by equation

$$P(X_{ij}=1|\theta,\beta_{i0},\ldots,\beta_{im_i})=\frac{\exp(j\theta+\beta_{ij})}{\sum_{l=0}^{m_i}\exp(l\theta+\beta_{il})}, \tag{2.1}$$

where $X_{ij}=1$ if S chooses response category C_{ij} of item I_i, and $X_{ij}=0$ otherwise; θ denotes the latent trait parameter of S, and β_{ij} an 'attractiveness' parameter of response category C_{ij} of item I_i, $j=0,\ldots,m_i$. (The parameters β_{ij} are often replaced by sums of certain 'threshold' parameters τ_{ij}, however, this reparameterization is of no advantage here. Under either parameterization, some normalization conditions are needed to render the parameters unique, however, the considerations in the present paper are independent of whatever normalization has been chosen.)

Since we are concerned with measuring change, we have to specify the model also for another time point T_2 at which the latent trait parameter of S may have changed from θ to $\theta+\eta$, where η denotes the amount of change in S between T_1 and T_2. The PCM for time point T_2 then becomes

$$P(X_{ij}=1|\theta,\eta,\beta_{i0},\ldots,\beta_{im_i})=\frac{\exp(j(\theta+\eta)+\beta_{ij})}{\sum_{l=0}^{m_i}\exp(l(\theta+\eta)+\beta_{il})}\,. \tag{2.2}$$

As is usual in IRT, the item responses are assumed to be 'locally independent', implying stochastic independence of S's responses as long as S's parameters θ and η are constant. (This does not preclude non-zero correlations between items in the population $\mathcal{P}$ of testees, though, because then θ and η may vary between persons.) Local independence will be postulated throughout this paper.

Under the PCM, the raw score r on a test with k items is $r=\sum_{i=1}^{k}\sum_{j=0}^{m_i}jx_{ij}$. This comprises, as a special case, the 'number right score' of intelligence tests with dichotomous items.

The questions as to how the parameters of the model can be made unique by normalization conditions and by which methods the normalized parameters can be estimated empirically have been treated sufficiently in psychometric literature, cf. Andersen (1995) or Fischer and Ponocny (1995); the interested reader is referred to these sources and the references therein. The preferred approach to estimation is the 'conditional maximum likelihood' (CML) method which capitalizes on the existence of a nontrivial conditional probability (or likelihood) of the data, given the testees'

raw scores; it is a function of the item × category parameters β_{ij}, but is independent of the person parameters θ. A CML program for the estimation of the parameters of an RM, an RSM, or a PCM is described in Fischer and Ponocny-Seliger (1998). (This WINDOWS program is available from ProGAMMA, Groningen.)

The conditional likelihood approach, based on (2.1) and (2.2), will also be used for the estimation of the change parameter η and for tests of hypotheses about η. It is easy to verify that the conditional likelihood of S's responses is independent of the chosen normalization of the PCM parameters β_{ij}. This independence is the reason why we need not bother about the normalization; it suffices to assume that parameters β_{ij} are given by which the PCM (2.1) and (2.2) is defined.

3 Some notation

The central idea underlying the present measurement of change and the testing of hypotheses about change within a PCM is the following: the argument in the exponential function in (2.2), $j(\theta+\eta)+\beta_{ij}$, can be rewritten as $j\theta+(\beta_{ij}+j\eta)=j\theta+\beta^*_{ij}$, where the β^*_{ij} are new item × category parameters of item I_i. The latter are sometimes designated as 'virtual' item parameters since no items with these parameters exist in reality. Employing this reparameterization implies that the person parameter θ remains constant, while change is projected into the virtual item parameters. The advantage of this is that formulae (2.1) and (2.2) can jointly be considered *one* PCM for a person S with a constant person parameter θ and with item × category parameters β_{ij} (for items I_i from the pretest, which henceforth will be denoted $\mathcal{I}_1$) and β^*_{lj} (for items I_l from the posttest, henceforth denoted $\mathcal{I}_2$). In other words, we may consider the pretest $\mathcal{I}_1$, given at time point T_1, and the posttest $\mathcal{I}_2$, given at time point T_2, jointly as one test of length k, comprising h items $I_i \in \mathcal{I}_1$ and $k-h$ items $I_l \in \mathcal{I}_2$.

For formal convenience, we shall replace the parameters β_{ij} by transformed parameters $\epsilon_{ij}=\exp(\beta_{ij})$, θ by $\xi=\exp(\theta)$, and η by $\delta=\exp(\eta)$, so that equations (2.1) and (2.2) have to be rewritten as

$$P(X_{ij}=1|\xi,\epsilon_{i0},\ldots,\epsilon_{im_i})=\frac{\xi^j\epsilon_{ij}}{\sum_{l=0}^{m_i}\xi^l\epsilon_{il}} \tag{3.1}$$

and

$$P(X_{ij}=1|\xi,\delta,\epsilon_{i0},\ldots,\epsilon_{im_i})=\frac{\xi^j\delta^j\epsilon_{ij}}{\sum_{l=0}^{m_i}\xi^l\delta^l\epsilon_{il}}. \tag{3.2}$$

Denote the maximum or 'perfect' scores attainable in $\mathcal{I}_1$ and $\mathcal{I}_2$ as $s_1=\sum_{i=1}^{h}m_i$ and $s_2=\sum_{i=h+1}^{k}m_i$, respectively, and let all cases be excluded where the total score $r=r_1+r_2$ is zero or s_1+s_2 (i.e. where r_1 and r_2 are fully determined as soon as r is given). The notation is summarized in Table 2. Notice that the two subsets of items $\mathcal{I}_1$ and $\mathcal{I}_2$ may be disjoint, or overlapping, or even identical. Therefore, our approach allows for many different testing designs with two testing occasions.

Table 2: Notation for a repeated measurement design with two time points.

Occasions	T_1	T_2
Subtests	$\mathcal{I}_1$	$\mathcal{I}_2$
Items	$\overbrace{I_1, ..., I_h}$	$\overbrace{I_{h+1}, ..., I_k}$
Responses	$\underbrace{\mathbf{x}_1, ..., \mathbf{x}_h}$	$\underbrace{\mathbf{x}_{h+1}, ..., \mathbf{x}_k}$
Response Patterns	$\mathbf{X}_1$	$\mathbf{X}_2$
Numbers of Items	h	$k-h$
Raw Scores	r_1	$r_2 = r - r_1$
Perfect Scores	s_1	s_2
Item × Category Parameters of the Real Items	$\underbrace{\epsilon_{10}, ..., \epsilon_{hm_h}}$	$\underbrace{\epsilon_{h+1,0}, ..., \epsilon_{km_k}}$
Parameter Vector	$\boldsymbol{\epsilon}$	
Item × Category Parameters of the Virtual Items	$\epsilon_{10}, ..., \epsilon_{hm_h}$	$\delta^0\epsilon_{h+1,0}, ..., \delta^{m_k}\epsilon_{km_k}$

4 Measurement of change

It is well-known that in the PCM the conditional probability of a testee's response pattern $\mathbf{X}$, given his/her raw score r, is independent of his/her person parameter ξ (or θ) (cf. Andersen, 1995). This probability is obtained by dividing the unconditional probability of $\mathbf{X}$ by the unconditional probability of observing any response pattern $\mathbf{Y}$ compatible with the given raw score r, the denominator being a certain combinatorial expression resulting from the summation over all patterns $\mathbf{Y}$ which yield r. With the present parameterization, this conditional probability is

$$P(\mathbf{X}|r) = \frac{(\prod_{i=1}^{k}\prod_{j=0}^{m_i}\epsilon_{ij}^{x_{ij}})\delta^{r_2}}{\gamma_r(\mathcal{I}_1, \mathcal{I}_2, \delta)}\ , \tag{4.1}$$

where $\gamma_r(\mathcal{I}_1, \mathcal{I}_2, \delta) = \gamma_r(\epsilon_{10}, \ldots, \epsilon_{hm_h}; \delta^0\epsilon_{h+1,0}, \ldots, \delta^{m_k}\epsilon_{km_k})$ denotes the said combinatorial expression; the latter is

$$\gamma_r(\mathcal{I}_1; \mathcal{I}_2, \delta) = \sum_{\mathbf{Y}|r}\Big(\prod_{i=1}^{h}\prod_{j=0}^{m_i}\epsilon_{ij}^{y_{ij}}\Big)\Big(\prod_{i=h+1}^{k}\prod_{j=0}^{m_i}(\delta^j\epsilon_{ij})^{y_{ij}}\Big), \tag{4.2}$$

the sum over $\mathbf{Y}$ meaning summation over all patterns $\mathbf{Y}$ that are compatible with r (cf. Andersen, 1995, p. 277).

Denoting that part of response pattern $\mathbf{Y}$ which pertains to $\mathcal{I}_1$, as $\mathbf{Y}_1$, and the rest, which pertains to $\mathcal{I}_2$, as $\mathbf{Y}_2$, with respective raw scores r_1 and r_2 satisfying $r_1 + r_2 = r$, it is seen that

$$\gamma_r(\mathcal{I}_1, \mathcal{I}_2, \delta) = \sum_{\substack{r_1, r_2 \\ r_1+r_2=r}} \Big(\sum_{\mathbf{Y}_1|r_1} \prod_{i=1}^{h} \prod_{j=0}^{m_i} \epsilon_{ij}^{y_{ij}} \Big) \Big(\sum_{\mathbf{Y}_2|r_2} \prod_{i=h+1}^{k} \prod_{j=0}^{m_i} \epsilon_{ij}^{y_{ij}} \Big) \delta^{r_2}, \tag{4.3}$$

where now the summation is taken over all r_1 and r_2 sufficing $r_1 + r_2 = r$. This again can be written more elegantly as

$$\gamma_r(\mathcal{I}_1, \mathcal{I}_2, \delta) = \sum_{l=a}^{b} \gamma_{r-l}(\mathcal{I}_1) \gamma_l(\mathcal{I}_2) \delta^l. \tag{4.4}$$

The limits of this summation, a and b, are determined as follows: evidently, for the summation index l the constraints $l \geq 0$, $l \geq r - s_1$, $l \leq s_2$, and $l \leq r$ must hold. These constraints are immediately seen to imply

$$a = \max(0, r - s_1), \tag{4.5}$$

$$b = \min(r, s_2). \tag{4.6}$$

It is now clear that

$$p_\eta(r_2) := P(r_1, r_2|r) = P(r_2|r) = P(r_2 - r_1|r) = \frac{\gamma_{r-r_2}(\mathcal{I}_1) \gamma_{r_2}(\mathcal{I}_2) \delta^{r_2}}{\gamma_r(\mathcal{I}_1, \mathcal{I}_2, \delta)}. \tag{4.7}$$

Note that these probabilities represent the conditional distribution of the gain score variable $R_2 - R_1$, which is identical to the conditional distribution of R_2; it depends on only one parameter, δ (or, equivalently, η), provided that the item $\times$ category parameters are known. It is easy to verify that this distribution is a one-parametric exponential family. As we shall see, this has a number of interesting and quite beneficial implications. (On exponential families, see Barndorff-Nielsen, 1978, or Andersen, 1980.)

Now it is straightforward to determine a CML estimator of the change parameter δ (or, equivalently, of η) by maximizing (4.7) in terms of δ, which yields an estimator $\hat{\eta}$ as our measure of change in testee S. Clearly, to do so the item $\times$ category parameters β_{ij} of the test must have been previously estimated from a sufficiently large sample of testees. This person sample must stem from a population $\mathcal{P}$ in which the PCM holds true, and the PCM must also hold for the present testee S, of course; but the sample need not be random or representative of the population, nor is it necessary to assume that S be sampled randomly from $\mathcal{P}$. This is due to the fact that the conditional estimator of the parameter vector $\boldsymbol{\epsilon}$ (or, equivalently, $\boldsymbol{\beta}$) is consistent irrespective of the distribution of the θ parameters in $\mathcal{P}$ (except for certain degenerate distributions; see Andersen, 1973).

Maximization of (4.7) with respect to δ can be done by taking the derivative of its logarithm and putting it to zero. The result of this, however, can directly be obtained also from a well-known theorem from the theory of exponential families: the ML (here: CML) estimator is obtained by equating the observed realization of

the statistic, r_2, with its conditional expectation, $E(R_2|r)$, which yields the CML estimation equation

$$r_2 = \sum_{l=a}^{b} p_\eta(l)l \;= \sum_{l=a}^{b} \frac{\gamma_{r-l}(\mathcal{I}_1)\gamma_l(\mathcal{I}_2)\delta^l l}{\gamma_r(\mathcal{I}_1,\mathcal{I}_2,\delta)} \;. \tag{4.8}$$

If (4.8) is solved for $\hat{\delta}$, an estimate $\hat{\eta} = \ln\hat{\delta}$ obtains, i.e. a measure of S's amount of change on the latent scale θ. Moreover, if it should be possible to show that the PCM, under suitable assumptions about the item universe, yields measures of the latent trait θ on an *interval* scale (as was done by Fischer, 1995b, p. 21, at least for the Rasch Model), it would follow that the estimator $\hat{\eta}$ is unique except for a multiplicative scaling constant, i.e. lies on a *ratio scale*. This would be is a very important property of change measurement because, e.g., in clinical psychology it would become possible to make statements like 'the amount of improvement (on the latent trait measured by the test scale) of testee S_a under treatment A is twice as large as that of testee S_b under treatment B.

Technical questions as to how to compute the functions γ numerically and how to solve the estimation equation (4.8) cannot be discussed within the scope of the present paper; the interested reader is referred to Fischer (2001). Here it suffices to mention that very efficient recursions are known for the γ (see Andersen, 1972, 1995; Fischer and Ponocny, 1994, 1995) and that (4.8) is most conveniently solved by means of a simple bisection method (see Fischer, 2001).

From well-known properties of exponential families it follows that, if the sufficient statistic R_2 is in the interior of the support [a,b], a finite CML estimator $\hat{\eta}$, i.e. a finite solution of (4.8), exists and is unique. Therefore, our procedure for the measurement of change will yield a unique result except for certain boundary cases. What happens with the root of (4.8) when $r_2 = a$ or $r_2 = b$ is seen in the overview given in Table 3 (on the proof, see Fischer, 2001).

5 Clopper-Pearson confidence intervals and some hypothesis tests

In many typical cases, an increase of the person parameter is expected, e.g., of abilities under training, therapy, or growth; or a decrease is expected, e.g., of abilities due to aging or of personality disorders under therapy. Therefore, it is of interest to test the H_0: $\eta = 0$ against the one-sided H_1: $\eta > 0$ or H_1: $\eta < 0$, respectively. Considering the conditional distribution (4.7), it is obvious that the tail probability (5.1),

$$P(R_2 \geq r_2|\boldsymbol{\beta},\eta_0,r) = \sum_{l=r_2}^{b} p_{\eta_0}(l), \tag{5.1}$$

Table 3: Existence and uniqueness of a CML estimator of η and of finite limits of Clopper-Pearson confidence intervals.

	Estimator		
R_2	$\hat{\eta}$	η_*	η^*
a	$-\infty$	$-\infty$	ext.
$a+1$	ext.	ext.	ext.
...	...	...	...
...	...	...	...
$b-1$	ext.	ext.	ext.
b	$+\infty$	ext.	$+\infty$

Note: The entry 'ext.' means that a finite solution exists and is unique. The results concerning lower limits η_* and upper limits η^* of Clopper-Pearson confidence intervals hold equally for the one-sided and two-sided cases.

can be used to test the H_0: $\eta_0 = 0$ against the H_1: $\eta > 0$, namely, to reject H_0 at significance level α if $P(R_2 \geq r_2|\boldsymbol{\beta}, \eta_0, r) \leq \alpha$. (The case H_1: $\eta < 1$ is analogous.)

An equivalent approach is to compute a one-sided so-called Clopper-Pearson confidence interval (η_*, ∞) for the change parameter η, using again the conditional distribution (4.7). The lower limit η_* is obtained by solving

$$P(R_2 \geq r_2|\boldsymbol{\beta}, \eta_*, r) = \alpha \tag{5.2}$$

with respect to η_* (cf. Santner and Duffy, 1989). For this purpose, again a bisection method turned out to be quite efficient and easy to apply, see Fischer (2001). (The case H_1: $\eta < 0$ is analogous.)

The Clopper-Pearson confidence intervals can also be interpreted as significance tests of the null-hypothesis of no change, H_0: $\eta = 0$. The H_0 is rejected if the point $\eta = 0$ falls outside the respective confidence interval. Such tests are generally conservative, like the underlying confidence intervals.

These simple but conservative confidence intervals and related hypothesis tests have several attractive properties. First, they are based on the exact conditional distribution of R_2, given the total observed score r, and hence do not require any asymptotic approximations. 'Asymptotic' in this connection means 'for $k \to \infty$', which would be an unrealistic assumption because most test scales have very limited length. Second, they are quite analogous to elementary methods, e.g., for the binomial distribution, described in textbooks of elementary statistics. Third, as will be illustrated by the example below, they are more attractive, from the point of view of an applied researcher, than optimal confidence intervals and most powerful tests, because the latter involve an additional element of randomness (see Section 7 below).

In the case of a two-sided alternative hypothesis, H_1: $\eta \neq 0$, a 'naive' – but simple

Table 4: Numbers of categories and parameters of the 14 self-assertiveness items, and the design of the pretest and posttest.

	m_i	β_{i0}	β_{i1}	β_{i2}	β_{i3}	Prt.	Pot.
1	1	0.0	1.72			1	0
2	1	0.0	0.87			0	1
3	1	0.0	-0.84			0	1
4	1	0.0	-1.40			1	0
5	2	0.0	0.97	1.09		0	1
6	2	0.0	0.89	0.65		1	0
7	2	0.0	0.36	-0.64		0	1
8	2	0.0	-0.33	-1.53		1	0
9	2	0.0	-0.61	-2.30		0	1
10	2	0.0	-0.83	-2.56		1	0
11	3	0.0	1.22	2.01	4.10	1	1
12	3	0.0	-0.14	-0.42	-1.93	1	0
13	3	0.0	-0.95	-1.96	-3.04	0	1
14	3	0.0	-1.48	-2.77	-6.58	1	1

Note: The items are arbitrarily ordered first by their numbers of categories, second by their (easiness) parameters β_{i1}. For normalization, all $\beta_{i0} = 0$ and $\sum_i \beta_{i1} = 0$. The item design is described by design vectors with elements 1 or 0, depending on whether an item is or is not contained in the pretest or posttest, respectively. The abbreviation 'Prt.' means 'Pretest', 'Pot.' means 'Posttest'. The maximum score of either test is 17.

– method is to apply exactly the same procedures as under the one-sided alternative hypotheses, however, with $\alpha/2$ at either tail of the conditional distribution. Such 'equi-tailed' intervals are analogous to many methods in elementary statistics, but in general there is no logical justification for splitting α into $\alpha/2$ at each tail; this is correct only in symmetric distributions. The present conditional distribution (4.7) is symmetric if $\mathcal{I}_1$ and $\mathcal{I}_2$ are identical item sets, but is non-symmetric in most other situations.

One and two-sided Clopper-Pearson intervals and significance tests for test scales conforming to the RM have already been implemented in the software LPCM-Win 1.0 by Fischer and Ponocny-Seliger (1998).

We conclude this section with results on the uniqueness of the lower limit, η_*, of a one-sided Clopper-Pearson confidence interval (η_*, ∞) and of the upper limit, η^*, of the one-sided interval $(-\infty, \eta^*)$. They are summarized in Table 3. (The proofs can again be found in Fischer, 2001). These results also hold for equi-tailed two-sided confidence intervals with $\alpha/2$ instead of α.

Table 5: Effect parameter estimates $\hat{\eta}$, significances, and confidence intervals (Clopper-Pearson and Randomized) for one-sided and two-sided alternative hypotheses at significance level $\alpha = .05$ for all combinations of r_1 and r_2 with total score $r = 19$.

			One-Sided H_1: $\eta > 0$				Two-Sided H_1: $\eta \neq 0$					
			Clopp.-Pear.		Randomiz.		Clopper-Pearson			Randomized		
r_1	r_2	$\hat{\eta}$	P	η_*	P	η_*	P	η_*	η^*	P	η_*	η^*
17	2	-∞	1.000	-∞	1.000	-∞	0.000	-∞	-3.55	0.000	-∞	-∞
16	3	-5.07	1.000	-8.59	1.000	-7.20	0.000	-9.31	-2.45	0.000	-∞	-3.17
15	4	-3.83	1.000	-6.18	1.000	-5.45	0.000	-6.60	-1.63	0.000	-7.76	-2.33
14	5	-2.93	1.000	-4.90	1.000	-4.27	0.002	-5.24	-0.92	0.001	-4.77	-0.99
13	6	-2.16	0.999	-3.96	0.996	-3.54	0.020	-4.25	-0.27	0.014	-3.89	-0.39
12	7	-1.48	0.990	-3.16	0.956	-2.99	0.130	-3.43	0.33	0.040	-3.42	-0.08
11	8	-0.85	0.935	-2.45	0.865	-2.17	0.466	-2.70	0.90	0.437	-2.15	0.89
10	9	-0.26	0.767	-1.79	0.677	-1.47	1.000	-2.04	1.46	0.781	-1.81	1.33
9	10	0.30	0.486	-1.19	0.389	-0.91	0.971	-1.43	2.03	0.918	-1.42	1.61
8	11	0.86	0.215	-0.62	0.180	-0.26	0.429	-0.86	2.64	0.365	-0.80	2.28
7	12	1.44	0.061	-0.07	0.020	-0.01	0.122	-0.30	3.33	0.048	0.01	3.28
6	13	2.08	0.010	0.49	0.004	0.64	0.020	0.26	4.13	0.003	0.83	4.24
5	14	2.82	0.001	1.10	0.001	1.57	0.002	0.87	5.12	0.000	1.42	5.39
4	15	3.71	0.000	1.85	0.000	1.94	0.000	1.54	6.49	0.000	1.97	7.39
3	16	4.96	0.000	2.61	0.000	2.80	0.000	2.33	9.21	0.000	2.66	∞
2	17	∞	0.000	3.78	0.000	4.65	0.000	3.45	∞	0.000	∞	∞

6 An example

Suppose scales for the measurement of 'Self-Assertiveness', each composed of a subset of a total of 14 items which conform to the PCM, are given to patients before and after a psychotherapy that aims at promoting self-confidence. Items 1 to 4 are dichotomous, items 5 to 10 have three categories each, and items 11 to 14 four categories each. The pretest is composed of items 1, 4, 6, 8, 10, 11, 12, 14; the posttest of items 2, 3, 5, 7, 9, 11, 13, 14. The maximum scores of pretest and posttest are both 17. The parameters and the test design are shown in Table 4.

For a demonstration, we arbitrarily select all score combinations r_1 and r_2 with total score $r = r_1 + r_2 = 19$. The CML estimates $\hat{\eta}$, their significances by formula (5.1), and Clopper-Pearson confidence intervals, based on one-sided and two-sided alternative hypotheses, are given in Table 5. (The randomized confidence intervals and related significances will be explained in Section 7.) Comparing the significance levels P, for all positive changes $\hat{\eta} > 0$ (i.e. for $r_2 = 10, 11, \ldots, 17$), under the two-sided H_1: $\eta \neq 0$ with those under the one-sided H_1: $\eta > 0$ shows that the latter are smaller (as they indeed should be): the one-sided tests are more powerful. For

Table 6: Significances on the basis of formula (5.1) under the one-sided H_1: $\eta > 0$ (left half of the table) and, with $\alpha/2$, under the two-sided H_1: $\eta \neq 0$ (right half of the table) for the self-assertiveness scale.

	1			1	
r_1/r_2	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7	r_1/r_2	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7	r_1/r_2
0	. s s S T T T T T T T T T T	0	. s S S T T T T T T T T T	0
1	. s S S T T T T T T T T	1	s s S T T T T T T T T	1
2	s s S S T T T T T T T	2	. s s S T T T T T T T	2
3	. s s S S T T T T T T	3	. s s S T T T T T T	3
4	. s s S S T T T T T	4	. s s S T T T T T	4
5	. s s S S T T T T	5	.	. s s S S T T T T	5
6	. s s S T T T	6	s	s s S S T T	6
7	. s s S T T	7	s s .	. s S T T	7
8	. s s S T	8	S s s .	. s S T	8
9	. s S T	9	T S s s .	. s T	9
10	. s S	10	T T S s s	s S	10
11	s S	11	T T T S s s	. s	11
12	. s	12	T T T T S s .	s	12
13	s	13	T T T T T S s .	.	13
14	.	14	T T T T T S S s s		14
15		15	T T T T T T S S s s .		15
16		16	T T T T T T T T S S s s		16
17		17	T T T T T T T T T T S S s s		17
	1			1	
	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7		0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7		

Note: The leftmost, middle, and rightmost columns give the pretest scores r_1, the top and bottom rows the posttest scores r_2. The interior of the table represents the levels of significance: blanks denote nonsignificance, '.' denotes significance at $\alpha = .10$, 's' at $\alpha = .05$, 'S' at $\alpha = .01$, and 'T' at $\alpha = .001$.

negative changes $\hat{\eta} < 0$ (i.e. for $r_2 = 9, 8, \ldots, 2$), it is clear that no significances are possible under the one-sided alternative H_1: $\eta > 0$. Under the two-sided H_1: $\eta \neq 0$, however, the scores $r_2 = 6, 5, \ldots, 2$ indicate a significant deterioration of Self-Assertiveness (at $\alpha = .05$). This is also reflected in the corresponding confidence intervals.

The significance levels for all score combinations $r_1 \times r_2$ are summarized in the form of Table 6 for H_1: $\eta > 0$ and for H_1: $\eta \neq 0$. Such tables are easy to apply even for researchers who are not familiar with statistical or psychometric methods since it suffices to look up the significance of the respective combination of scores r_1 and r_2.

Two aspects have to be kept in mind. First, such a table is valid only for fixed pretests and posttests, $\mathcal{I}_1$ and $\mathcal{I}_2$. If different testee groups are presented with

different item sets, one table has to be computed for each group. In individualized (computerized adaptive) testing, however, such tables become impractical; then it is necessary to compute the significance level for each individual separately. Second, equal gain scores d may be differently significant depending on the respective r_1 and r_2 values. As Table 6 (left half) shows, e.g., for H_1: $\eta > 0$, a gain score of $d = 7 - 2 = 5$ is significant at the $\alpha = .05$ level, while the gain score $d = 13 - 8 = 5$ of the same size is only significant at $\alpha = .10$. (This alone suffices to illustrate why it would not make sense to assign a single indicator of measurement precision like, e.g., reliability, to all values of the variable $D = R_2 - R_1$.) Comparing this significance level, on the other hand, with that of the same scores under the two-sided H_1: $\eta \neq 0$, it is seen that in the latter case there is no significance; this is understandable because the two-sided test is less powerful.

7 Uniformly most accurate confidence intervals and uniformly most powerful unbiased tests

Although the Clopper-Pearson confidence intervals and the related probabilities (5.1) are intuitively appealing and easy to understand for researchers in psychology or education, their disadvantage is that they lead to conservative decisions: the H_0 of no change is sometimes retained where a more powerful procedure might lead to its rejection. Fortunately, a detailed theory of Uniformly Most Powerful Unbiased (UMPU) tests and of the related Uniformly Most Accurate (UMA) confidence intervals (on this terminology, see Mood, Graybill, and Boes, 1974) for one-parametric families of exponential distributions can be found in statistical literature (see, for instance, Lehmann, 1986, or Witting, 1985). The idea underlying the construction of a UMPU test of the H_0 of no change and of a UMA confidence interval for η is to transform the discrete random variable R_2 into a continuous variable $W = R_2 + U$, where R_2 and U are stochastically independent and U has a rectangular distribution on $[0, 1)$. This transformation is called 'randomization' of R, and the resulting procedures are denoted 'randomized tests' and 'randomized confidence intervals', respectively. Variable W has a continuous distribution function A that is strictly monotone decreasing in η,

$$A_\eta(w) = \sum_{l=a}^{[w]-1} p_\eta(l) + (w - [w])p_\eta([w]) \ , \tag{7.1}$$

where $[w]$ denotes the largest integer $\leq w$ and the sum is defined to be zero if $[w] - 1 < a$. It is important to understand that, although a random 'error' variable U is added to R_2, no loss of information is incurred because, if w is given, r_2 can still be uniquely inferred by taking the nearest integer $\leq w$.

The continuous variable W can now be used for the construction of optimal tests and confidence intervals. The lower limit of the one-sided UMA confidence interval

Table 7: Existence and uniqueness of finite limits of UMA confidence intervals.

	H_1: $\eta > 0$		H_1: $\eta < 0$		H_1: $\eta \neq 0$	
	η_*	η^*	η_*	η^*	η_*	η^*
$W \leq a + \alpha$	$-\infty$	$+\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
$a + \alpha < W \leq a + 1 - \alpha$	$-\infty$	$+\infty$	$-\infty$	ext.	$-\infty$	ext.
$a + 1 - \alpha < W \leq a + 2 - \alpha$	ext.	$+\infty$	$-\infty$	ext.	$-\infty$	ext.
$a + 2 - \alpha < W < b - 1 + \alpha$	ext.	$+\infty$	$-\infty$	ext.	ext.	ext.
$b - 1 + \alpha \leq W < b + \alpha$	ext.	$+\infty$	$-\infty$	ext.	ext.	$+\infty$
$b + \alpha \leq W < b + 1 - \alpha$	ext.	$+\infty$	$-\infty$	$+\infty$	ext.	$+\infty$
$b + 1 - \alpha \leq W$	$+\infty$	$+\infty$	$-\infty$	$+\infty$	$+\infty$	$+\infty$

Note: The entry "ext." means that a finite solution exists and is unique. In the case of two-sided alternative hypotheses, however, uniqueness holds only if $[c_1] < [c_2]$.

(η_*, ∞) under H_1: $\eta > 0$ is obtained by solving (7.2) for η_*,

$$A_{\eta_*}(w) = 1 - \alpha. \tag{7.2}$$

Similarly, the upper limit η^* of the one-sided confidence interval $(-\infty, \eta^*)$ under H_1: $\eta < 0$ is obtained by solving (7.3) for η^*,

$$A_{\eta^*}(w) = \alpha. \tag{7.3}$$

The numerical problems involved are again conveniently overcome by means of a simple bisection method.

The question as to the uniqueness of limits η_* and η^* and of their behavior when R_2 assumes values at or near the boundaries of the interval $[a, b]$ is more complicated than in the case of Clopper-Pearson intervals because it depends not only on the realization of R_2 but also on the random component U and on the significance level α. The results are seen in Table 7. (The proof can be found in Fischer, 2001.)

Equivalent to the construction of a UMA confidence interval is a UMPU test of H_0 against H_1: $\eta > 0$ or against H_1: $\eta < 0$: instead of solving (7.2) or (7.3), with given α, for η_* or η^*, we now insert $\eta = 0$ in (7.2) or (7.3), respectively, and compute α, which is the significance level of the test of H_0: $\eta = 0$.

For the dichotomous RM, UMPU tests have already been suggested by Klauer (1991a), Liou and Chang (1992), Liou (1993), and Ponocny (2000). Some tests are implemented, e.g., in the software T-Rasch (Ponocny and Ponocny-Seliger, 1999). Fischer (2001) generalized the results of Klauer (1991a) to the entire family of models defined by the PCM.

The effect of randomization on the one-sided confidence intervals (η_*, ∞) and on one-sided tests of H_0: $\eta = 0$ is illustrated in Tables 5, 6, and 8 (left half); the latter table gives the significance levels for score combinations r_1 and r_2 for the example

Table 8: Significances based on randomized tests under H_1: $\eta > 0$ (left half of the table) and under H_1: $\eta \neq 0$ (right half of the table) for the self assertiveness scale.

	1		1	
r_1/r_2	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7	r_1/r_2	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7	r_1/r_2
0	s . S T S T T T T T T T T T	0	s s s S T T T T T T T T T	0
1	. s S S S T T T T T T T T	1	. s T S T T T T T T T	1
2	. . s S S S T T T T T T T	2	. s s S T T T T T T	2
3	. s S S T T T T T T T	3	s s S S S T T T T T	3
4	s s s S T T T T T T	4	. . s S S S T T T T	4
5	. . S s S S T T T T	5	. s S s S T T T T	5
6	s s s S S T T T	6	s . . s s S S T T T	6
7	. . s s S S T T	7	S s . s s s S S T	7
8	. s s S T T	8	S S s s S S S T	8
9	s s S T	9	T S S s s s S T	9
10	. s s S	10	T T S S s . . s s S	10
11	. S T	11	T T T T S s s s . S	11
12	s S	12	T T T S S s s s . S	12
13	s T	13	T T T T T T s s . s	13
14	s	14	T T T T T S T s s . . .	14
15	S	15	T T T T T T T S S s s	15
16		16	T T T T T T T T S S s S .	16
17		17	T T T T T T T T T T S S S s . .	17
	1		1	
	0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7		0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7	

Note: The leftmost, middle, and rightmost columns give the pretest scores r_1, the top and bottom rows the posttest scores r_2. The interior of the table represents the levels of significance: blanks denote nonsignificance, '.' denotes significance at $\alpha = .10$, 's' at $\alpha = .05$, 'S' at $\alpha = .01$, and 'T' at $\alpha = .001$.

of Section 6 based on randomized UMPU tests. Comparing first the corresponding columns in Table 5 shows that the randomized one-sided confidence intervals are always better and the significance levels lower than in the Clopper-Pearson case; this is as expected. Turning then to the contours of significance levels, a comparison of Tables 6 (left half) and 8 (left half) reveals that the contours in the latter table are more 'ragged'. Moreover, Table 6 possesses the property of 'double monotonicity', meaning that significance levels increase or decrease monotonically both row- and columnwise. This certainly is in accordance with what empirical researchers would expect. Table 8, on the other hand, partly violates this property and thus – at least at first sight – seems to contain some contradictory entries. This is not so, however: from a theoretical perspective, monotonicity has to hold only in all directions parallel to the secondary diagonal of the table, that is, *within* each conditional distribution with fixed total score r. As the reader can easily check, Table 8 indeed satisfies

this monotonicity requirement. Anyway, the advantage of the randomized tests in Table 8, as compared to those in Table 6, lies in their greater power to reject the H_0: $\eta = 0$.

From the point of view of an applied researcher, the UMPU randomized tests have one serious disadvantage, though: applying such a test to two testees with the same total and gain scores, or performing the test twice for the same person (keeping the scores fixed), may occasionally produce different results (different significance levels). This is probably not in accordance with most researchers' expectations about methodology. The statistician, however, can easily explain this random behavior of the test as formally correct and logical: if, for instance, a simple Student's t-test is carried out twice to test the H_0 of equal means of two (sub)populations, to repeat the test means to draw another pair of independent samples and to compute the test statistic anew, which may obviously lead to a result that contradicts a previous one. In the same way, repeating the significance test of the difference of two person parameters in the present case implies observing independently new test scores (which should be possible by virtue of the local independence assumption) and subsequently drawing a new random number u from the uniform distribution on [0,1). Clearly, total and posttest scores may be observed on the second occasion that differ from those previously observed, and a different realization u of the random variable U will also occur. Therefore, the outcome of the test in terms of a significance level may be different. Even repeating only the computation of the significance test involves sampling a new u, which may already suffice to produce a change in the significance level. Nevertheless, the present chance dependence of the significance test of a gain score might present a greater problem to the empirical researcher than using the less powerful Clopper-Pearson tail probabilities.

Turning now to the case of a two-sided alternative hypothesis H_1: $\eta \neq 0$, the method of randomization can be applied as before, but it entails more complicated procedures. According to the theory of statistical tests in one-parametric exponential families, a UMPU test of H_0: $\eta = 0$ can be written as a function ϕ of the observation $w = r_2 + u$,

$$\phi(w) = \begin{cases} 1 & \text{for } w \leq c_1 \text{ or } w \geq c_2, \\ 0 & \text{otherwise,} \end{cases} \tag{7.4}$$

where $W = R_2 + U$ is the randomized score variable, $\phi = 1$ means rejecting and $\phi = 0$ retaining the H_0, and c_1, c_2 are certain cutoff scores. The latter can be determined from the system

$$E_{\eta_0}[\phi(W)] = \alpha, \tag{7.5}$$

$$E_{\eta_0}[R_2\phi(W)] = \alpha E_{\eta_0}(R_2), \tag{7.6}$$

with $\eta_0 = 0$ under the present H_0 of no change (cf., e.g., Witting, 1985, Chap. 2, who describes in much detail the procedure for the binomial distribution; see also Klauer, 1991a, and Klauer, 1991b, who applied the same to the RM).

Evaluation of the expectations in (7.5) and (7.6) yields the system

$$A_{\eta_0}(c_1) + B_{\eta_0}(c_2) = \alpha, \tag{7.7}$$

$$\tilde{A}_{\eta_0}(c_1) + \tilde{B}_{\eta_0}(c_2) = \alpha, \tag{7.8}$$

where $A_\eta(c_1)$ is the distribution function (7.1) taken at $w = c_1$, and

$$B_\eta(c_2) = ([c_2] + 1 - c_2)p_\eta([c_2]) + \sum_{l=[c_2]+1}^{b} p_\eta(l), \tag{7.9}$$

$$\tilde{A}_\eta(c_1) = \frac{\sum_{l=a}^{[c_1]-1} p_\eta(l)l + (c_1 - [c_1])p_\eta([c_1])[c_1]}{\sum_{l=a}^{b} p_\eta(l)l}, \tag{7.10}$$

$$\tilde{B}_\eta(c_2) = \frac{([c_2] + 1 - c_2)p_\eta([c_2])[c_2] + \sum_{l=[c_2]+1}^{b} p_\eta(l)l}{\sum_{l=a}^{b} p_\eta(l)l}. \tag{7.11}$$

(Again, a sum where the lower summation limit is greater than the upper limit is defined to be zero.)

The system of equations (7.7) and (7.8) affords a basis for solving two equivalent problems, the construction of a UMA confidence interval (η_*, η^*) and the determination of the cutoff scores c_1 and c_2 for a two-sided UMPU test of H_0: $\eta = \eta_0$. (On this equivalence, see Mood, Graybill, and Boes, 1974, p. 464). The technical problems of how to solve the system (7.7)-(7.8) and the questions related to the uniqueness of the results are treated in Fischer (2001). The conditions for the uniqueness of (η_*, η^*) are summarized in Table 7. As in the case of one-sided alternative hypotheses, the uniqueness of a solution depends on r_2, on the realization u of the random component U, and on the chosen significance level α. Therefore, it may again happen that, for a given pair of scores r and r_2 together with a realization u, a unique solution is found, but that, upon recalculation with another u, one of the limits η_* and η^* diverges to $-\infty$ or ∞.

8 The example continued

To give a rounded picture of the procedures outlined in this paper, UMA confidence intervals and UMPU significance levels under the two-sided H_1: $\eta \neq 0$ are also presented in Tables 5 and Table 8 (right half). To see the difference between the Clopper-Pearson approach and the randomized confidence intervals and tests, compare the corresponding columns in Table 5: for almost all score combinations, the two-sided randomized confidence intervals are narrower, but near the extreme score combinations $r_1 = 2$, $r_2 = 17$ and $r_1 = 17$, $r_2 = 2$, one or both limits of the UMA confidence interval tend to diverge. The reader may further wish to compare the significance levels of the UMPU two-sided tests (Table 8, right half) with those of the Clopper-Pearson procedure (Table 6, right half). As is clear on theoretical grounds, the former are always smaller than the latter, that is, the UMPU tests are more powerful.

9 Conclusion

Treating the problem of the change of raw scores between two time points within the framework of a family of Rasch models (RM, RSM, PCM) leads to a powerful arsenal of tools: simple Clopper-Pearson confidence intervals and related significances as well as UMA confidence intervals and related UMPU hypothesis tests based on randomized scores. All of them are 'exact' in the sense that they are grounded on the exact conditional distribution of score differences, given the observed total score r, which implies that no asymptotic approximations are required. They allow one to compute tables of the significance of score differences which are easy to apply in practice. All these methods hold for any selection of subtests $\mathcal{I}_1$ and $\mathcal{I}_2$ from a given item pool that conforms to the postulated model (PCM). By the way, the same methods and tables also apply to the statistical assessment of score differences between two different testees.

The present author advocates that such tables should routinely become part of test manuals of published tests. Clearly, this is only possible if pretest and posttest are fixed. An example of such a use of the methods outlined in this paper can be found in the Manual for Raven's Progressive Matrices and Vocabulary Scales (Appendix 3 by Fischer and Prieler, 2000).

From all this it is evident that the 'precision' of change measurements can be evaluated on the basis of confidence intervals of the change parameter η and of related significance levels. Such a characterization of precision is somewhat more complex than the traditional one because more than just one number (i.e. reliability) is required. On the other hand, the present approach does not suffer from the paradoxes and weaknesses inherent in the concept of reliability.

Further extensions of the present methods where individual significances of change are integrated into one significance of change for a group of testees (e.g., in clinical research with small groups) are immediate but cannot be treated within the scope of the present paper.

References

[1] Andersen, E.B. (1972): The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, **34**, 42–54.

[2] Andersen, E.B. (1973): *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygiejnisk Vorlag.

[3] Andersen, E.B. (1980): *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

[4] Andersen, E.B. (1995): Polytomous Rasch models and their estimation. In G.H. Fischer and I.W. Molenaar (Eds.): *Rasch Models: Foundations, Recent Developments, and Applications* 271–291. New York: Springer-Verlag.

[5] Andrich, D. (1978): A rating formulation for ordered response categories. *Psychometrika*, **43**, 561–573.

[6] Barndorff-Nielsen, O. (1978): *Information and Exponential Families in Statistical Theory.* New York: Wiley.

[7] Cronbach, L.J. and Furby, L. (1970): How should we measure change, or should we? *Psychological Bulletin*, **74**, 68–80.

[8] Fischer, G.H. (1995a): Some neglected problems in IRT. *Psychometrika,* **60**, 459–487.

[9] Fischer, G.H. (1995b): Derivations of the Rasch model. In G.H. Fischer and I.W. Molenaar (Eds.): *Rasch Models: Foundations, Recent Developments, and Applications*, 15–38. New York: Springer-Verlag.

[10] Fischer, G.H. (2001): Gain scores revisited under an Irt perspective. In A. Boomsma, M.A.J. van Duijn, and T.A.B Snijders (Eds.): *Essays on Item Response Theory* 43–68. *Lecture Notes in Statistics*, **157**. New York: Springer-Verlag.

[11] Fischer, G.H. and Ponocny, I. (1994): An extension of the partial credit model with an application to the measurement of change. *Psychometrika,* **59**, 177–192.

[12] Fischer, G.H. and Ponocny, I. (1995): Extended rating scale and partial credit models for assessing change. In G.H. Fischer and I.W. Molenaar (Eds.): *Rasch models: Foundations, Recent Developments, and Applications* 351–370. New York: Springer-Verlag.

[13] Fischer, G.H. and Ponocny-Seliger, E. (1998): *Structural Rasch Modeling. Handbook of the Usage of LPCM-WIN 1.0.* Groningen, NL: ProGAMMA.

[14] Fischer, G.H. and Prieler, J.A. (2000): An IRT methodology for the assessment of change. In J. Raven, J.C. Raven, and J.H. Court (Eds.): *Standard Progressive Matrices. Raven Manual: Section 3*, 113–121. Oxford: Oxford Psychologists Press.

[15] Harris, C.W. (1963): *Problems in Measuring Change.* Madison, WI: University of Wisconsin Press.

[16] Klauer, K.C. (1991a): An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika,* **56**, 213–228.

[17] Klauer, K.C. (1991b): Exact and best confidence intervals for the ability parameter of the Rasch model. *Psychometrika,* **56**, 535–547.

[18] Lehmann, E.L. (1986): *Testing Statistical Hypotheses* (2nd ed.). New York: J. Wiley.

[19] Liou, M. (1993): Exact person tests for assessing model-data fit in the Rasch model. *Applied Psychological Measurement,* **17**, 187–195.

[20] Liou, M. and Chang, C.-H. (1992): Constructing the exact significance level for a person fit statistic. *Psychometrika,* **47**, 169-181.

[21] Lord, F.M. and Novick, M.R. (1968): *Statistical Theories of Mental Test Scores.* Reading: MA: Addison-Wesley.

[22] Masters, G.N. (1982): A Rasch model for partial credit scoring. *Psychometrika,* **47**, 149–174.

[23] Mood, A.M., Graybill, F.A., and Boes, D.C. (1974): *Introduction to the Theory of Statistics.* Singapore: McGraw-Hill.

[24] Ponocny, I. (2000): Exact person fit indexes for the Rasch model for arbitrary alternatives. *Psychometrika,* **65**, 29–42.

[25] Ponocny, I. and Ponocny-Seliger, E. (1999): *T-Rasch 1.0.* Groningen, NL: ProGAMMA.

[26] Rasch, G. (1960): *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

[27] Rasch, G. (1965): *Statistisk Seminar.* [Statistical seminar.] (Notes taken by J. Stene.) Copenhagen: Department of Statistics, University of Copenhagen.

[28] Santner, T.J. and Duffy, D.E. (1989): *The Statistical Analysis of Discrete Data.* New York: Springer-Verlag.

[29] Willett, J.B. (1989): Some results on reliability for the longitudinal measurement of change: implications for the design of studies of individual growth. *Educational and Psychological Measurement,* **49**, 587–602.

[30] Williams, R.H. and Zimmerman, D.W. (1996): Are simple gain scores obsolete? *Applied Psychological Measurement,* **20**, 59–69.

[31] Witting, H. (1985): *Mathematische Statistik I.* [Mathematical Statistics, Vol. I.] Stuttgart: Teubner.