A Mixture Proportional Hazards Model With Random Effects for Response Times in Tests

Educational and Psychological
Measurement

2016, Vol. 76(4) 562–586

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415598347

epm.sagepub.com

Jochen Ranger¹ and Jörg-Tobias Kuhn²

Abstract

In this article, a new model for test response times is proposed that combines latent class analysis and the proportional hazards model with random effects in a similar vein as the mixture factor model. The model assumes the existence of different latent classes. In each latent class, the response times are distributed according to a class-specific proportional hazards model. The class-specific proportional hazards models relate the response times of each subject to his or her work pace, which is considered as a random effect. The latent class extension of the proportional hazards model allows for differences in response strategies between subjects. The differences can be captured in the hazard functions, which trace the progress individuals make over time when working on an item. The model can be calibrated with marginal maximum likelihood estimation. The fit of the model can either be assessed with information criteria or with a test of model fit. In a simulation study, the performance of the proposed approaches to model calibration and model evaluation is investigated. Finally, the model is used for a real data set.

Keywords

response time, proportional hazards model, latent class analysis

¹Martin Luther University Halle-Wittenberg, Halle, Germany

²University of Münster, Münster, Germany

Corresponding Author:

Jochen Ranger, Martin Luther University Halle-Wittenberg, Department of Psychology, D-06099 Halle (Saale), Germany.

Email: jochen.ranger@psych.uni-halle.de

Introduction

Psychological assessment deals with the measurement of human characteristics such as ability, characteristic traits, attitudes, or preferences. The process of measurement is thereby based on the responses to standardized stimuli, which are usually short statements or isolated microproblems bundled together in a test. These responses are then mapped to the continuum of the characteristic that is supposed to underlie the test performance. The mapping requires a measurement model that describes the relation between the observable responses and the unobservable characteristic, the so-called latent trait. Standard measurement models are the logistic models that were originally developed for achievement tests (Baker & Kim, 2004). The application of a particular measurement model requires that the relation between the responses and the trait is the same for all individuals. This implies that all individuals respond in the same way, or more specifically, that the items trigger the same response processes in all subjects. This assumption, however, might be invalid in some cases. There is accumulating evidence that individuals differ in the way they process information and arrive at a response.

With respect to achievement tests research findings indicate the existence of subgroups that differ qualitatively in their way of responding. In tests with little consequences for the test takers there is sometimes a subgroup of individuals that guesses rapidly while the majority of the test takers strive to solve the items. This problem occurs in low-stakes testing where some subjects lack the motivation to make any effort (Bolt, Cohen, & Wollack, 2002; Meyer, 2010; Schnipke, 1999; Wise & DeMars, 2006). Qualitatively different ways of responding, however, can have more profound reasons. In diagnostic classification models, the tests are designed to distinguish subgroups of test takers who hold qualitative different forms of misconceptions and thus rely on different mental operations when responding; see, for example, Bradshaw and Templin (2014). Although misconceptions become manifest in the response, the way individuals apply concepts also influences the response time (Lasry, Watkins, Mazur, & Ibrahim, 2013). Sometimes, individuals even differ in a more fundamental way. The dual processing theory, which was developed in cognitive psychology, states that there are two modes of information processing, an automated mode and a controlled mode. Individuals switch from controlled processing to automated processing with practice (Goldhammer et al., 2014). Automated processing is supposed to be faster than controlled processing.

Evidence for the presence of subgroups of test takers who respond in a different way can also be found in personality tests and attitudinal scales. An important topic in psychological assessment is the problem that some individuals respond on some other basis than the specific item content. Response styles like acquiescence or extreme responding can be mentioned here as well as the motivation to respond in a social desirable way. Much effort has been spent on the identification of the affected subjects. In addition to the responses the response times are indicative of the response style, and several researchers managed to separate different groups of responders by means of a response time analysis (Holden & Kroner, 1992; Hsu, Santelli, & Hsu,

1989; McIntyre, 2011; Schneider & Hübner, 1980). However, not all subjects that respond differently should be classified as aberrant responders. For attitudinal scales it has been shown that individuals can be classified into two groups on the basis whether they take a rapid peripheral route of information processing or a more central one (Mayerl, 2013). In the same line one can distinguish strong and weak attitudes, a distinction that refers to the availability of the attitudes when completing the test. Attitudes are either retrieved or generated when responding to the test items and it is the retrieved attitudes that are stable and predict behavior (Bassili, 1996; Fazio, 1995; Mulligan, Grant, Mockabee, & Monson, 2003). Retrieving attitudes is usually faster than generating them and it has been shown that the response time can be used to separate the two response processes. Lately, the existence of different classes of respondents has also been noticed in the field of conjoint measurement. Latent class models have been used successfully for choice data to identify unobserved consumer segments differing in preference structures and decision-making strategies (Jedidi & Kohli, 2005; Swait & Adamowicz, 2001).

Summing up, this short overview illustrates three points. First, there is much evidence that individuals do not only differ in the trait level but sometimes also in the general way to process the items of the test. This allows for a grouping of the subjects into several latent classes. The number and nature of these classes depend on the content of the test and the situation the assessment takes place. Considering the overview given above, however, very often it should be just two latent classes that have to be distinguished. Depending on the context this could be the classes of subjects that respond rapidly versus in a well-thought-out way, that use controlled versus automated processes, that answer honestly versus social desirably, or that take a peripheral route versus a central route of information processing.

Second, it is of diagnostic value to identify the presence of latent classes and to classify the subjects into the class they belong to. Each latent class is characterized by a different relationship between the response and the latent trait and therefore requires a class-specific measurement model to avoid wrong conclusions about the traits of the test takers. The problem of invalid trait inference however is only one aspect. The classification of subjects into different classes might be of diagnostic value itself in case one is able to identify rapid guessers, individuals with strong attitudes, or different consumer segments. Knowledge of the class membership allows for a more detailed characterization of the individuals that is known to serve as a moderator variable of the trait behavior relationship; see, for example, the publications about attitude strength or social desirableness.

And third, the response time might be a good indicator of the latent class membership. Very often, the time course of the response process differs significantly for different ways to approach an item. In the case of automated versus controlled processing or the central versus peripheral route it is the response time and not so much the response that distinguishes the two solution processes. Rapid guessing becomes apparent not only in a large amount of wrong responses but also in short response times. And even in the case of social desirability it is the response time that

has been discussed as the primary indicator of this response style. Not surprisingly, it is the response time that has very often been used to identify subgroups of test takers responding in a different way; see the references cited above. This is similar to experimental psychology, where the response time is considered as the key quantity in order to analyze the response process of subjects.

Consequently, in this article, we propose a model for response times in tests that allows for different response processes. The model combines two well-known approaches to response time modeling, latent class analysis, and the proportional hazards model with random effects. The latent class model assumes that individuals can be grouped into different latent classes with distinct, class-specific response time models. The class-specific response time models are assumed to be proportional hazards models with random effects that account for individual differences in work pace. This approach parallels the mixture factor analysis model of Yung (1997) when the standard factor model is replaced by the proportional hazards model. Here, the proportional hazards model was chosen for two reasons. First, it is the most popular model for event times in biometrics because of its flexibility and semiparametric approach. Second, as highlighted by Wenger and Gibson (2004) and Ranger and Ortner (2013) the model has a precise psychological interpretation in terms of an information accumulation process and can be used to estimate the rate of information acquisition of the subjects. This is an advantage over more statistical models that are not closely related to elements of the solution process and lack a psychological interpretation.

The outline of the article is as follows. First, the proportional hazards model with random effects is described and an interpretation in terms of an information accumulation process is given. Then, the model is extended to a latent class version. Model estimation and approaches to model inference are described in a separate section. In this section, we also address ways to test the fit of the model. Finally, the model is applied to a real data set.

The Proportional Hazards Model and Its Interpretation

The proportional hazards model of Cox (1972) is the most popular model for the analysis of event times in statistics. It has been introduced in psychology by Douglas, Kosorok, and Chewing (1999). Ranger and Ortner (2012) adapted the proportional hazards model to response times in tests by including a random effect representing the work pace of an individual. Similarly the proportional hazards model was used by Wang, Chang, and Douglas (2013) and Wang, Fan, Chang, and Douglas (2013) as a special variant of the more general linear transformation model. A proportional hazards model with crossed random effects has been proposed by Loeys, Legrand, Schettino, and Pourtois (2014).

The basic quantity in the proportional hazards model is the hazard function. The hazard function is a function of time and is defined as

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < t + \Delta t | T \geq t]}{\Delta t}, \quad (1)$$

where $\lambda(t)$ can be any continuous function that is positive and has an infinite integral over $[0, \infty)$. In the present context, the hazard function can be interpreted as the tendency to respond in the next moment provided that an individual has not responded before (Klein & Moeschberger, 1997). Because of its close relationship to a response in the next instant it can be interpreted as the momentary rate at which individuals process information. When a large amount of information is acquired per instant an individual has a high instantaneous tendency to respond. Consequently, the hazard function is a precise description of the response process that conveys more information than any summary statistic (Aalen & Gjessing, 2001; Wenger & Gibson, 2004). More details concerning this interpretation and a sound mathematical derivation why the hazard function represents the rate of information acquisition are given in Appendix A.

The hazard function determines the response time distribution as there is a one-to-one relation between the response time density and the hazard function. The assumption of different response time distributions over subjects and items implies that for each item g and each subject i there is a distinct hazard function $\lambda_{gi}(t)$. The proportional hazards model is a regression-like model that attributes the individual differences in the hazard function to predictive characteristics of the subjects. The basic assumption thereby is the claim that the hazard function in each item and subject can be composed of an item-specific baseline hazard function $\lambda_{B_g}(t)$ common to all subjects and the work pace θ_i of an individual via the multiplicative model

$$\lambda_{gi}(t; \theta_i) = \lambda_{B_g}(t) \exp(\beta_g \theta_i). \quad (2)$$

The baseline hazard function $\lambda_{B_g}(t)$ determines the response time distribution of an average individual with $\theta_i = 0$. It reflects the general time demand of an item. Although the baseline hazard function is a function and not a scalar its role is similar to the role of the intercept in a regression model or the role of the item location in a logistic item response model. The baseline hazard function is associated with item characteristics like the item difficulty as it is harder to process information in difficult or controversial items (Ferrando & Demestre, 2007; Klein Entink, Kuhn, Hornke, & Fox, 2009). However, similar to the item location in item response models the quantity is simply estimated and not related to item characteristics (although this would be possible). Item parameter β_g is a regression coefficient that controls the influence of work pace on the hazard function. Its function is similar to that of the loadings in factor analysis. Note that although the latent work pace is considered as a random variate we will not use different symbols for the random variate and its realization for ease of readability.

Using the well-known relationship between the hazard function and the density function—see, for example, Klein and Moeschberger (1997)—the density of the response time distribution follows from Equation (2) as

$$f_g(t|\theta_i) = \lambda_{B_g}(t) \exp(\beta_g \theta) \exp[-\exp(\beta_g \theta) \Lambda_{B_g}(t)], \quad (3)$$

where $\Lambda_{B_g}(t) = \int_0^t \lambda_{B_g}(u) du$ denotes the cumulative baseline hazard function. The proportional hazards model will be extended to latent classes next.

A Mixture Proportional Hazards Model

The fusion of latent trait and latent class models has a long tradition in psychometrics. The basic idea of such models is the combination of continuous latent variables in form of the latent trait with discrete latent variables in form of a latent class membership (Vermunt, 2008). Several such models have been proposed for the responses from psychological tests. Mixture item response models assume different item response models for several not directly observable subgroups of the test takers (Magidson & Vermunt, 2001; Rost, 1990). Mixture factor models postulate different factor models for likewise not directly observable subpopulations (Varriale & Vermunt, 2012; Yung, 1997). Recently, combinations of latent class models with latent trait models have also been suggested for response time data in tests, predominantly in the context of rapid guessing (Meyer, 2010). One could also model test response data with one of the latent class factor models suggested in biometrics for event data (Almansa, Vermunt, Forero, & Alonso, 2014; Asparouhov, Masyn, & Muthen, 2006). However, instead of using one of the modeling approaches proposed in the past that either are based on the normal distribution or that can be used for discrete time data only, we propose a latent class latent trait model that is based on the proportional hazards model for continuous time in this article.

When combining latent trait and latent class analysis, two different strategies can be pursued, the k -mixtures of common latent trait models versus the k -mixtures of heterogeneous latent trait models (Yung, 1997). In the first approach, a common latent trait model applies to all subgroups of the population. The distribution of the latent trait however is different in the latent classes such that the marginal distribution of the latent trait across the latent classes follows a mixture of distributions. In the second approach, each subgroup can be distinguished by a group-specific latent trait model. The latent trait, however, is assumed to be standard normally distributed in each subgroup. It is the second approach that is followed here. More specifically, we assume that there are different subgroups of test takers that differ in their way of responding. Each way of responding can be characterized by a distinct rate of information acquisition, that is, by a distinct baseline hazard function. For sake of simplicity, we assume just two subgroups. Two latent subgroups might be the first choice as very often there are two fundamental different ways of responding; see the introduction for examples. Each test taker can be assigned to one of the two subgroups according to his/her way of responding. Note, that this assumption implies that individuals do not change their way of responding during the test. The group membership of the individuals cannot be observed directly. Therefore, the subgroups are denoted as latent classes.

Represent by $c_i \in \{1, 2\}$ the latent class membership of an individual. The proportional hazards model can be extended to latent classes by the assumption that conditional on the latent class the distribution of the response time in an item can be related to the work pace θ_i of an individual according to the class-specific proportional hazards model:

$$\lambda_{gci}(t; \theta_i, c_i) = \lambda_{B_{gc}}(t) \exp(\beta_{gc} \theta_i). \quad (4)$$

Here, $\lambda_{B_{gc}}(t)$ is a class-specific baseline hazard function and β_{gc} a class-specific regression coefficient. In case the latent class membership and the work pace of an individual are the only systematic influences on the response process, one can assume that the response times in all items of the test are conditionally independent. This is a standard assumption typically made in latent trait and latent class modeling. With the conditional independence assumption, the distribution of the response times $\mathbf{t} = (t_1, \dots, t_G)$ in the G items of the test factors to

$$f(\mathbf{t}|\theta_i, c_i) = \prod_{g=1}^G f_g(t_g|\theta_i, c_i), \quad (5)$$

where $f_g(t_g|\theta_i, c_i)$ denotes the density function given in Equation (3) with the class-specific hazard function given in Equation (4). The marginal distribution of the response times in each class follows from Equation (5) after integration over distribution $f(\theta|c_i)$, that is, the distribution of the latent trait θ in the latent class the individual belongs to, as

$$f(\mathbf{t}|c_i) = \int \prod_{g=1}^G f_g(t_g|\theta, c_i) f(\theta|c_i) d\theta. \quad (6)$$

In accordance with the standard assumption typically made in item response and structural equation models the distribution $f(\theta|c_i)$ is assumed to be a normal distribution. As the latent classes are assumed to be mutually exclusive, the overall response time distribution follows as the mixture of the class-specific marginal response time distributions given in Equation (6)

$$f(\mathbf{t}) = \pi_1 \int \prod_{g=1}^G f_g(t_g|\theta, c=1) f(\theta|c=1) d\theta + \pi_2 \int \prod_{g=1}^G f_g(t_g|\theta, c=2) f(\theta|c=2) d\theta, \quad (7)$$

where π_1 and π_2 denote the mixing proportions, that is, the probability of belonging to one of the two latent classes. Note that in case the latent class of the subjects was known, the model would simplify to a stratified survival model for multivariate response times, a so-called stratified frailty model.

In the present form, the model is not identified. Scale and location differences of the class-specific latent trait distributions $f(\theta|c = 1)$ and $f(\theta|c = 2)$ can be compensated by the class-specific regression coefficients β_{gc} and the class-specific baseline hazard functions $\lambda_{B_{gc}}(t)$. Because the latent traits are not observable, some restrictions have to be imposed in order to identify the model. Similar to the standard factor model or the mixture of heterogeneous latent trait models (Yung, 1997) we assume that the latent trait follows a standard normal distribution in each class. This fixes the location and the scale of the latent trait and identifies the model.

Equation (7) is similar to the k-mixtures of heterogeneous factor models of Yung (1997) with the exception that the class-specific distributions of the response times are not the multivariate normal distributions implied by the standard factor model but follow from the proportional hazards model with random effects. The model also bears resemblance to the multivariate multilevel model for continuous survival times described in Asparouhov et al. (2006). The main advantage of the proportional hazards framework is its flexibility and its interpretability in terms of an information accumulation process. The class-specific baseline hazard functions allow for a precise assessment of how the subgroups differ with respect to information processing. In the previous section, only two latent classes have been considered. This is often the first choice as a model should be as parsimonious as possible and there are good reasons to assume just two fundamental different ways of responding. This could imply a group guessing rapidly versus a group responding adequately or a group responding honestly versus a group concerned with impression management. Theoretically, the approach could be extended to more latent groups. The number of latent classes is only limited by the ability to calibrate the model. However, it might be the better strategy to prefer a simpler model over a more complex one that is hard to fit to the data (Hastie, Tibshirani, & Friedman, 2009).

Implementation of the Model and Model Calibration

One of the key quantities of the model is the baseline hazard function. It can be interpreted as the rate at which an average individual in a latent class acquires information. Therefore, the class-specific baseline hazard functions should be the most informative quantity to compare the different solution processes in the two latent classes. The hazard function also determines the response time distribution. When the hazard function is constant the response times are distributed according to an exponential distribution. Linear or quadratic functions result in Weibull distributed response times. When implementing the model, strong assumptions about the form of the class-specific baseline hazard functions should be avoided. In fact, the primary motivation to use the proportional hazards model is its potential to record the process of information acquisition in a flexible way. The required flexibility can be achieved by approximating the class-specific baseline hazard functions $\lambda_{B_{gc}}(t)$ in a piecewise manner, using different functions for disjoint segments of the time axis. With a

sufficiently large number of segments such an approximation can assume virtually any shape.

A standard approach to piecewise continuous modeling consists in the usage of polynomial spline functions (Hastie et al., 2009). Polynomial spline functions are polynomials on each segment with continuous derivatives of a certain order across the whole range of their domain. Polynomial spline functions are usually generated by a linear combination of basis functions. In general, the B-spline basis is used. The resulting linear combination however is not necessarily positive as it is required by the baseline hazard function. Although the positivity of polynomial splines can be guaranteed by imposing restrictions on the weights of the linear combination (Etezadi-Amoli & Ciampi, 1987) this approach is not easy to implement. It is easier to approximate the baseline hazard function with an approach proposed by Ramsay (1988). The approach is based on a different basis, the so-called M-spline basis. The M-spline basis functions are likewise combined via a linear combination to a piecewise continuous polynomial. Due to characteristics of the basis functions, the positivity of the weights guarantees the positivity of the piecewise polynomial such that no complicated restrictions are necessary. A disadvantage, however, is the fact that the M-spline basis with positive weights is somewhat restricted as it does not cover the whole space of positive piecewise polynomials. This is usually not a disadvantage in practice as the M-Spline usually suffices for most shapes of the baseline hazard function; see Cai, Lin, and Wang (2011) and Younes and Lachin (1997) for an application of M-splines in response time modeling.

Because of space limitations, the approach based on M-splines will only be sketched here. More details can be found in Appendix B. Piecewise modeling of the class-specific baseline hazard functions requires that the time axis is divided into disjoint segments at knots that define the segment borders. In each segment, the hazard function is modeled by a distinct polynomial of degree m . The segment-specific polynomials, however, are subject to restrictions as the resulting function is required to be continuous with $m - 1$ continuous derivatives. A piecewise quadratic function for example is a quadratic function in each segment with continuity up to the first derivative. The continuity requirements guarantee the smoothness of the function. The piecewise function can be generated by a linear combination of basis functions, the M-spline basis. With q segments $K = q + m$ basis functions have to be defined for a piecewise polynomial of degree m . These basis functions will be denoted by $M_{k,m}(t)$ in the following with $k = 1, \dots, K$. The basis functions can be generated by divided differences of truncated power functions; see Appendix B for a recursive definition. More important than their generation are the properties of the M-spline basis functions. Each basis function is itself a piecewise polynomial. The basis functions are all positive over at most $m + 1$ segments and 0 elsewhere and integrate to 1. Figure 1 depicts the five basis functions needed for a piecewise quadratic function with a continuous first derivative based on three segments separated at 3.33 and 6.66.

The piecewise polynomial function is generated by a linear combination of the M-spline basis functions. The weights of the linear combination have to be restricted

Figure 1. M-spline basis functions for a piecewise quadratic function and three segments defined by segment borders at 3.33 and 6.66. An exemplified linear combination of the basis functions with positive weights yields the bold line that represents a potential approximation of the baseline hazard function.

to be positive. This restriction guarantees the positivity of the resulting polynomial because all M-spline basis functions are positive. This allows the approximation of the baseline hazard functions as follows. In a first step, several time points have to be chosen for each item that divides the time axis into segments. Then, the M-spline basis functions have to be determined for each set of time points. The class-specific baseline hazard function $\lambda_{B_{gc}}(t)$ in each item is approximated by $\lambda_{B_{gc}}(t) = \sum_k \alpha_{gck} M_{k,m}(t)$ where α_{gck} are class-specific item parameters that determine the shape of the piecewise polynomial in item g and class c . See Figure 1 where a baseline hazard function is generated with the weights $\alpha_{gc1} = 0.34$, $\alpha_{gc2} = 2.63$, $\alpha_{gc3} = 0.99$, $\alpha_{gc4} = 0.49$, and $\alpha_{gc5} = 0.20$. Note that the resulting linear combination is a positive, quadratic function in each segment and possesses a continuous first derivative.

The application of the model requires the estimation of the unknown parameters, that is, the regression coefficients from the proportional hazards model, the mixing

proportions, and the weights needed for the M-spline approximation of the class-specific baseline hazard functions. The knots that define the segments are regarded as given. Usually, one places the knots at equidistant quantiles or spreads them uniformly over the time continuum. The unknown parameters can be estimated with marginal maximum likelihood estimation as it is standard practice in latent trait modeling. Marginal maximum likelihood estimation is based on the marginal response time distribution that can be obtained as follows. First, one has to replace the baseline hazard functions in Equation (4) with the M-spline approximation $\lambda_{B_{gc}}(t) = \sum_k \alpha_{gck} M_{k,m}(t)$. Substituting the approximation in Equation (3) and Equation (5) gives the distribution of the response times in the test conditional on the latent class membership and the individual work pace. Each class-specific conditional distribution has to be integrated over work pace as illustrated in Equation (6). The integrated class-specific response time distributions constitute the marginal response time distribution according to Equation (7). It is this distribution marginal maximum likelihood estimation is based on.

Denote by vector $\gamma = [\pi, \beta, \alpha]$ the vector of the unknown model parameters consisting of the mixing proportions $\pi = [\pi_1, \pi_2]$, the regression coefficients $\beta = [\beta_{11}, \beta_{12}, \dots, \beta_{G1}, \beta_{G2}]$ and the weights $\alpha = [\alpha_{11}, \alpha_{12}, \dots, \alpha_{G1}, \alpha_{G2}]$ for the G items and the two latent classes. Having observed the response time pattern $t = [t_1, \dots, t_N]$ of the N test takers in a calibration sample, the marginal log-likelihood function can be defined as

$$ML(\gamma; t) = \sum_{i=1}^N \log[f(t_i)], \quad (8)$$

where $f(t_i)$ is the distribution given in Equation (7). Marginal maximum likelihood estimation consists in determining the values of the parameters that maximize the marginal log-likelihood function. The standard approach to the maximization of the marginal log-likelihood function in mixture models is the expectation maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) although any alternative optimizer could be used as well; see McLachlan and Krishnan (1997) for more details concerning the implementation of the EM algorithm in mixture models. Mixture models are known to have many local maxima. Therefore, it is common practice to estimate a model several times with different initial values.

Analysis of Model Fit

Before any substantial interpretation of the research findings can be attempted two questions have to be investigated: the correct number of latent classes and the overall fit of the model to the data. Testing for the correct number of latent classes is not straightforward in mixture models. The standard approach of comparing nested models with different numbers of latent classes via a likelihood ratio test is invalid as the mixture parameters are on the boundary of the parameter space and the model is not

identified under the null hypothesis (Lo, Mendell, & Rubin, 2001). In latent class analysis or mixture item response models one usually uses information criteria to determine the correct number of latent classes. Popular information criteria are the Akaike information criterion (AIC) of Akaike (1992) and the Bayes information criterion (BIC) of Schwarz (1978). Although their application is also not justified in presence of boundary value parameters (Hughes & King, 2003) these measures or variants thereof usually perform well in practice, at least in larger samples (Bulteel, Wilderjans, Tuerlinckx, & Ceulemans, 2013; Henson, Reise, & Kim, 2007; Li, Cohen, Kim, & Cho, 2009; Nylund, Asparouhov, & Muthen, 2007). Therefore, in the present context we also advise the usage of information criteria for inferences about the number of the latent classes.

The overall fit of the model can be assessed with a test proposed by Ranger and Kuhn (2014). A test of model fit is usually superior to graphical checks in highly parameterized models. The test proposed by Ranger and Kuhn (2014) is a χ^2 -like test of goodness of fit similar to the χ^2 -test of Chernoff and Lehmann (1954). The test requires the binning of the response times at some prespecified time points $[d_0 = 0, d_1 = t_1, \dots, d_R = \infty]$. These time points define time slices for which the similarity of the observed and expected number of responses is assessed. The choice of the time points is somewhat arbitrary. They do not have to coincide with the knots used for the approximation of the baseline hazard function. Usually, one divides the time axis into about 10 equally spaced segments. Having selected the time points the observed numbers of responses $\mathbf{o}_g = [o_g(d_0, d_1), \dots, o_g(d_{R-1}, d_R)]^t$ given during each time slice have to be determined for all items $g = 1, \dots, G$. These quantities are then compared with the corresponding expected numbers of responses $\mathbf{e}_g = [e_g(d_0, d_1), \dots, e_g(d_{R-1}, d_R)]^t$ that are predicted for each time slice and item by the model and the estimated parameters. The expected number of responses follows from Equation (7) by an integration over time and a marginalization with respect to the response times of the other items. The similarity of the observed and expected numbers of responses are evaluated by means of the residuals $\mathbf{o}_g - \mathbf{e}_g$ that are squared and summed to a test statistic reflecting model fit. Depending on whether the fit of a single item or the fit of the whole test is assessed, the squared residuals of a single item or of all items are summed. As can be shown the sum of the squared residuals is approximately chi-square distributed. The degrees of freedom have to be corrected though in order to account for the complication that the same data set is used for model calibration and model evaluation; see Ranger and Kuhn (2014) for more details.

Simulation Study

In order to assess the proposed approach to model estimation a simulation study was conducted. This simulation study served two purposes, to assess the parameter recovery of the maximum likelihood estimator and to evaluate the proposed methods for inference. The simulation study was based on the response times in a test of 12

items. Two latent classes were assumed with mixture proportions of $\pi_1 = 0.75$ and $\pi_2 = 0.25$. In the first latent class, the response times followed a Weibull distribution with a shape parameter of 2. The baseline hazard function was $\lambda_{B_{g1}}(t) = 2 \cdot \exp(0.2) \cdot t$ in all items. The regression coefficients were $\beta_{1g} = 0.4$ in item $g = 1, \dots, 4$, $\beta_{1g} = 0.5$ in item $g = 5, \dots, 8$ and $\beta_{1g} = 0.6$ in item $g = 9, \dots, 12$. This setup implied class-specific expected response times of 0.82, 0.83, and 0.84 and variances of 0.22, 0.24, and 0.28 in the three item sets. The response times in the second latent class followed an exponential distribution. The baseline hazard function was $\lambda_{B_{g2}}(t) = \exp(0.2)$ in all items. The regression coefficients were $\beta_{2g} = 0.2$ in item $g = 1, \dots, 4$, $\beta_{2g} = 0.3$ in item $g = 5, \dots, 8$, and $\beta_{2g} = 0.4$ in item $g = 9, \dots, 12$. This setup implied class-specific expected response times of 0.84, 0.87, and 0.89 and variances of 0.75, 0.87, and 1.06 in the three item sets. Mixing both distributions with mixture proportions of $\pi_1 = 0.75$ and $\pi_2 = 0.25$ resulted in marginal expectations of 0.82, 0.83, and 0.85 and variances of 0.35, 0.40, and 0.47 in the three item sets. This simulation setup was considered to be rather demanding. The expected response times were almost the same in the two latent classes and the class-specific densities overlapped considerably. The marginal response time density was also unimodal and did not reveal the presence of two subgroups visually.

The simulation samples were obtained as follows. In a first step, the respondents were randomly assigned to one of the two latent classes according to the Bernoulli distribution and the mixture proportions. Then, the latent trait values were drawn from the standard normal distribution for every fictitious test taker in the sample. Finally, the response times were generated for each test taker by using the proportional hazards model of the latent class the test taker belonged to. In this way, 250 samples with 1,000 subjects and 250 samples with 5,000 subjects were generated.

Simulation Study: Parameter Recovery

In a first step, the model was fit to the simulation samples. The model was implemented as described above. The time axis was divided into three segments at the knots $\kappa_1 = 0.60$ and $\kappa_2 = 0.85$. The baseline hazard function was approximated by a positive piecewise quadratic function with continuous first derivative. When estimating the model, the response times were right and left censored at the time points $\kappa_0 = 0.35$ and $\kappa_3 = 1.20$. The response times were censored to reduce the influence of extreme observations and increase the robustness of the estimates (Ulrich & Miller, 1994). Censoring is easily accounted for in the proportional hazards model by a slight modification of the marginal likelihood function. Note that because of censoring the baseline hazard function can only be estimated between $\kappa_0 = 0.35$ and $\kappa_3 = 1.20$.

The integral over the normal distribution needed for marginal maximum likelihood estimation, see Equation (6), was approximated with Gauss Hermite Quadrature using 20 nodes. The marginal log-likelihood function was maximized

Table 1. Parameter Recovery of the Regression Coefficients.

N = 1,000							
Parameter	β_{g1}			β_{g2}			π_1
Item	1-4	5-8	9-12	1-4	5-8	9-12	—
Value	0.400	0.500	0.600	0.200	0.300	0.400	0.750
M	0.400	0.502	0.607	0.213	0.329	0.412	0.735
SE	0.070	0.072	0.079	0.167	0.205	0.230	0.042
CI	0.942	0.953	0.940	0.955	0.943	0.947	0.864
N = 5,000							
Parameter	β_{g1}			β_{g2}			π_1
Item	1-4	5-8	9-12	1-4	5-8	9-12	—
Value	0.400	0.500	0.600	0.200	0.300	0.400	0.750
M	0.400	0.501	0.603	0.202	0.303	0.405	0.744
SE	0.029	0.031	0.034	0.066	0.071	0.079	0.015
CI	0.951	0.947	0.944	0.954	0.957	0.952	0.940

Note. Results for β_{gc} and the mixture proportion π_1 as well as coverage frequencies of confidence intervals with a confidence level of $c = 0.95$ in the simulation study for different sample sizes N . Results based on 250 replications. Results are aggregated over items with the same item parameter. Value = True value; M = average estimate; SE = standard error of estimation, CI = relative frequency of coverage.

with the EM-algorithm. The algorithm was implemented in the statistical software environment R (R Development Core Team, 2009). The scripts are available from the authors on request. As mixture models usually possess several local maxima the performance of the estimator depends crucially on good initial values. The initial values were determined as follows. In a first step, the log response times were analyzed with the package mclust which estimates a finite mixture model based on the multivariate normal distribution (Fraley, Raftery, & Scrucca, 2014). A two-class solution was enforced and each test taker was assigned to one of the two latent classes. Then the total response time of each subject was used as a proxy for the latent trait. A provisional latent trait estimate was generated by transforming the total response time into a quantile of the standard normal distribution via quantile matching. Using the preliminary class memberships and the provisional latent traits of the subjects two proportional hazards model were estimated for every item and the two groups of respondents.

For each of the 500 (2 sample sizes $\times$ 250 replications) data sets the item parameters were estimated as described above. The EM-algorithm converged each time. Having estimated the item parameters, confidence intervals were calculated for the mixture proportion and the regression coefficients. Therefore, the standard error of the estimates was taken from the diagonal entries of the inverse of the observed Fisher information matrix. Then, Wald-type confidence intervals were calculated. The results can be found in Table 1 which contains the true value, the average

estimate, the standard error of estimation, and the relative coverage frequency of the confidence intervals for confidence level $c = 0.95$. Note that the results have been aggregated over items with the same item parameter (Items 1-4, 5-8, and 9-12).

The results in Table 1 imply that the maximum likelihood estimator performs well. The estimates of the parameters are virtually unbiased. However, the estimates are slightly imprecise in samples of 1,000 subjects. Especially the estimates of the regression coefficients in the second latent class have a large standard error. This might be due to the smaller mixing proportion ($\pi_2 = 0.25$) of this class. The validity of the confidence intervals depends strongly on the sample size. In samples of 1,000 subjects the coverage frequency of the intervals for the mixing proportion is lower than $c = 0.95$. The erratic performance is due to the sample distribution of the estimates which was skewed and far from the normal distribution. A likelihood ratio-based confidence interval might have performed better. The coverage probabilities of the regression coefficients are close to the confidence level though. In samples of 5,000 subjects all confidence intervals perform exactly as they should.

In addition to the recovery of the item parameters, we also investigated the recovery of the class-specific baseline hazard functions (momentary rate of information accumulation) and the cumulative baseline hazard functions (total accumulated information). In general, the functions could be recovered almost without bias. However, the standard error of estimation was rather large for the class-specific baseline hazard functions, especially in the class with the lower mixing proportion. This is not surprising as the hazard function is notoriously hard to estimate. More details concerning the estimation of the two quantities can be obtained from the authors on request.

Simulation Study: Model Fit

The performance of the information criteria and the test of model fit were evaluated in a second simulation study. Therefore, different simulation conditions were defined by varying the number of latent classes of the data-generating model. In the condition with two latent classes, the response times were generated as before, mixing the Weibull distribution with the exponential distribution. In the condition with one latent class, all response times were drawn from an exponential distribution. This exponential distribution was identical to the one of the first simulation condition. In the condition with three latent classes, the response times were generated by mixing three distributions. The first and second distribution were the Weibull and exponential distribution of the first simulation study. The third distribution was similar to the Weibull distribution of the first simulation condition with the exception that all response times were shifted (increased) by the fixed amount of 0.5. The mixing proportions were 0.25, 0.50, and 0.25. The resulting distribution was still unimodal. Again, 250 data sets of 1,000 and 5,000 subjects were generated for each simulation condition.

The data sets were analyzed as described above. First, the mixture model for two latent classes was fit to each of the $2 \times 3 \times 250$ data sets using the marginal

Table 2. Relative Frequency With Which a Model Was Chosen as the Best Approximation by the AIC and BIC Indices for Generating Models With Different Number of Latent Classes in Samples of 1,000 Subjects.

Classes	AIC			BIC		
	1	2	3	1	2	3
1	0.972	0.028	0.000	1.000	0.000	0.000
2	0.000	0.976	0.024	1.000	0.000	0.000
3	0.000	0.804	0.196	0.000	1.000	0.000

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion. Results based on 250 simulation samples.

maximum likelihood estimator. In addition, a modified version for one and three latent classes was fit to the data as well. However, only the data sets with 1,000 subjects were analyzed with the version for three latent classes as fitting the model was extremely time-consuming. Despite the fact that the models were misspecified in case the structure of the data set did not correspond to the assumed number of latent classes all estimators converged. Having estimated the parameters the AIC and BIC indices were calculated for all models and all data sets. The model with the lowest information criterion was chosen as the true one. The relative frequency of the choices is tabulated in Table 2 for the different simulation conditions and samples of 1,000 subjects. Note, that the diagonal in each subtable denotes the rate with which the number of latent classes was identified correctly.

The performance of the information criteria is suboptimal. The BIC fails at detecting the correct number of latent classes when two or more classes are present. This is somewhat surprising although it is known that the BIC sometimes performs poorly in more complex models (Bulteel et al., 2013). The AIC detects the presence of two latent classes but also fails at revealing the existence of three. The low detection rate in case of the third simulation condition is due to the good approximation of the marginal response time distribution by the two-class model. P-P plots that compared the correspondence of the true response time distribution with the one implied by the two-class model did not indicate the misspecification either. Given the variance of the Weibull distribution a shift of 0.5 appears negligible. As the AIC selects the model with the best trade-off between variance and bias it tends to favor the model with fewer parameters. Sample size adjusted versions of the criteria that reduce the penalty for the model complexity might improve the detection rate (Tofighi & Enders, 2007).

The performance of the proposed tests of model fit was analyzed as follows. In a first step, the two-class mixture model was fitted to each data set from the first simulation condition (two latent classes) and the third simulation condition (three latent classes). Then, the response times of every item were categorized into 10 time slices at 8 equally spread thresholds ranging from 0.4 to 1.10. With the categorized response

Table 3. Empirical Type I Error Rate (Condition LCA 2) and Power (Condition LCA 3) for Two Tests of Model Fit in Samples of 1,000 and 5,000 Subjects and Different Nominal Type-I Error Rates α .

Condition	Sample	α	Test: RT	Test: RT-G
LCA 2	1,000	0.10	0.090	0.092
		0.05	0.045	0.040
		0.01	0.009	0.012
	5,000	0.10	0.100	0.100
		0.05	0.056	0.040
		0.01	0.013	0.012
LCA 3	1,000	0.10	0.131	0.248
		0.05	0.073	0.124
		0.01	0.015	0.052
	5,000	0.10	0.185	0.612
		0.05	0.105	0.456
		0.01	0.023	0.200

Note. The tests evaluate the adequacy of a two-class model either itemwise (Test: RT) or globally (Test: RT-G). In Condition LCA 2, the data were based on two latent classes and in Condition LCA 3 on three. Results are based on 250 simulation samples. Tests for item fit are averaged over items.

times two tests of model fit were run. The first test was the test of item fit and compared the expected and observed numbers of responses in the time slices for each item separately. Such a test would be used during test development for item selection. The second test evaluated the congruence of observed and expected numbers of responses in all items jointly, thereby testing the overall fit of the model. Such a test is similar to the overall test of model fit in confirmatory factor analysis. The empirical rejection rates of the tests can be found in Table 3 for several nominal Type I error rates and the two simulation conditions. The reported relative frequencies are the empirical Type I error rate in the condition with two latent classes and the power of the test in the condition with three latent classes. Note, that the results from the item-specific version of the test have been averaged over the items.

As can be seen in Table 3, the tests adhere closely to the nominal Type I error rate. Note that for $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.10$ the empirical rejection rates are expected to be in a range of ± 0.01 , ± 0.03 , and ± 0.04 around their true value. The test has little power to detect the misspecification of the model in the three-class condition when the samples consist of 1,000 subjects. This finding parallels the results for the information criteria described above. The power increases in the samples of 5,000 subjects. Note that the global test assessing the fit of all items jointly is superior to the item-specific tests. This illustrates the limitation of graphical checks of model fit as these plots only assess the model fit itemwise. In addition to the proposed test of model fit we also inspected the performance of simple likelihood ratio tests comparing a two latent class version of the model with a three-latent class version. The performance of this test however was suboptimal and will not be reported.

Empirical Example

To demonstrate the usefulness of the model a real data set was analyzed. The data came from a mental calculation test given to second to fourth graders in a large-scale educational study. Here, only the first eight items that were given to all pupils were used. These items contained addition and subtraction problems that varied with respect to their difficulty. The items were constructed by systematically varying the number of digits (one vs. two) and the occurrence of a decade crossing (no decade crossing vs. decade crossing). The data set contained the responses and response times from 980 pupils.

The data were analyzed as described before. The response times were censored at roughly the q_{10} and q_{90} quantiles in each item. The time continuum between the censoring boundaries was partitioned into three segments by two knots. Then the proposed model was fit to the data. Several starting values were used and the best solution was chosen. Versions of the model with one and three latent classes were estimated as well. The AIC indices were $AIC_1 = 18153.99$, $AIC_2 = 1447.61$, and $AIC_3 = 17473.15$ for the one-, two-, and three-class version. This result suggested the version with two latent classes as the best model. The results of the fit tests however questioned the adequateness of the model. The overall test of model fit based on eight equally spaced cut points indicated a significant deviation of the observed and implied response time distribution ($\chi^2 = 40.70$, $df = 26.77$, $p = .04$). Inspecting the single items revealed that only the first item exhibited misfit ($\chi^2 = 18.15$, $df = 4.01$, $p < .01$). Therefore, this item was removed and the data analysis was repeated. Again, the version of the model with two latent classes was best ($AIC_1 = 17002.04$, $AIC_2 = 16418.34$, $AIC_3 = 16479.41$). Having removed the first item the overall fit of the model was excellent ($\chi^2 = 22.75$, $df = 22.80$, $p = .46$).

The two latent classes had mixing proportions of $\pi_1 = 0.66$ and $\pi_2 = 0.34$. Interestingly, although the subjects of the first latent class needed longer to respond (the average total testing time was 18.48 seconds vs. 12.62 seconds), there were little differences with respect to mathematical ability. The subjects from the first latent class scored 6.61 on average while the subjects from the second class scored 6.86. In a pretest of mathematical ability the differences between the two classes were also not remarkable (mean: 101.63, $SD = 14.4$ vs. mean: 105.63, $SD = 15.3$). However, the latent classes differed with respect to the age structure. In the first latent class, the pupils were mostly from the second and third grades (Grade 2: 0.41, Grade 3: 0.40, Grade 4: 0.19) whereas in the second latent class, the pupils came mostly from the third and fourth grades (Grade 2: 0.11, Grade 3: 0.47, Grade 4: 0.41). This may point to age-related differences in the calculation strategy or an automatization between the classes.

This interpretation was also supported by the estimates of the baseline hazard function. The shape of the functions differed considerably over the two latent classes. This points to qualitative differences in the solution process. Although the baseline hazard function cannot be estimated too precisely in small samples and should be interpreted with care, some patterns seemed to emerge. In the second latent class

there was a sharp peak around 1.3 seconds that appeared independently of the item difficulty. The hazard functions of the first latent class did not exhibit this sharp increase. Given that the second latent class consists of the older subjects one might conjecture that these subjects have switched from controlled to automated processing. Note that all items were rather easy and should be prone for automaticity.

Discussion

Latent trait models in general and item response models in particular are popular tools in psychological assessment. These models are based on the assumption that all individual differences are due to different levels of an underlying trait and therefore require that all individuals respond in the same way. However, sometimes there is also heterogeneity with respect to the response process. When this heterogeneity is ignored and a standard latent trait model is used as a measurement for the underlying latent trait the conclusions about the test takers are possibly wrong. The identification of subgroups that respond differently is important not only from a measurement perspective but also from a theoretical one and could improve test development and theory induction. Therefore, a model for the detection of different subgroups would be of practical value, especially in case the model provides some insight into how the groups differ. One candidate of such a model is the mixture proportional hazards model with random effects proposed in the article.

The model combines latent class analysis and the proportional hazards model and inherits the strengths of both approaches. The latent class analysis allows for subgroups differing with respect to the response process. The proportional hazards model provides a model for the information accumulation process in each subgroup. This allows for a detailed comparison of the different response processes via the subgroup-specific baseline hazard function. Although the model is for response times only, one could easily extend the model to a joint model for the responses and response times. One simply had to include a mixture item response model via the hierarchical framework of van der Linden (2007).

The model can be calibrated with marginal maximum likelihood estimation. The performance of the estimator is good, but the precise recovery of the item parameters requires a large sample. Large sample sizes are also needed for valid confidence intervals and a good performance of the proposed approaches to model inference. The correct number of latent classes can be recovered well in case the number is less than 3. Three latent classes are hardly detected, at least for the simulation conditions. This is partly due to the difficulty to fit the version of the model for three classes. Problems like several local minima of the marginal likelihood function are aggravated here. Furthermore, the model for two latent classes already provides a good approximation to the response time distribution. However, as the latent classes were not well separated, the simulation conditions were rather challenging. With a clearer separation of the latent classes, for example, when individuals either guess rapidly or answer conscientiously, the findings might be different.

The model was used for the analysis of an empirical data set. This analysis was supposed to be a demonstration of a prototypical application of the model. In this application pupils could be classified according to whether mathematical operations have already become automatized or not. But of course, the model has more potential applications. It can be used to detect rapid guessing or social desirable responding. It could be used for the identification of strong attitudes that are related more closely to behavior. The detection of differential item functioning is also a potential field of application. The usefulness of the model thereby stems from two sources: The model is flexible and avoids too strong assumptions about the response time distribution. Additionally, it provides a profound interpretation in psychological terms due to its relation to the process of information acquisition. Therefore, the model is not only a measurement model but also bears a resemblance to cognitive process models from experimental psychology. This is a new approach that might become popular in the future (van der Maas, Molenaar, Maris, Kievit, & Boersboom, 2011).

Appendix A

Interpretation of the Proportional Hazards Model

The proportional hazards model has an interpretation in terms of an information accumulation process. This interpretation can be traced back to Singpurwalla (2006); see Ranger and Ortner (2013) for an adaptation of this interpretation to the area of psychological tests. Due to space restrictions the justification of this interpretation will only be sketched here; we also simplify the original arguments slightly; see the references for more details. The core of the interpretation is the assumption that individuals acquire information until they are able to respond. The expected momentary rate at which individuals acquire information per instant is represented by the hazard function $\lambda_{gi}(t; \theta_i)$. The expected amount of information acquired up to time t is the integral of the momentary rate of information acquisition $\Lambda_{gi}(t; \theta_i) = \int_0^t \lambda_{gi}(u; \theta_i) du$. Individuals however differ from the expected rate of information acquisition. Similar to regression analysis, where individuals fluctuate around the regression line, we assume that the actual amount of acquired information of an individual with work pace θ_i deviates from the expected level of acquired information due to the effect of a residual influence ε_{gi} . This influence is supposed to represent all unsystematic influences on the response process not accounted for by work pace. The actual level of information $K_{gi}(t; \theta_i, \varepsilon_{gi})$ at time t is related to the expected level of information $\Lambda_{gi}(t; \theta_i)$ and the residual influence ε_{gi} via the log-linear model $\log[K_{gi}(t; \theta_i, \varepsilon_{gi})] = \log[\Lambda_{gi}(t; \theta_i)] + \varepsilon_{gi} = \log[\Lambda_{B_g}(t)] + \beta_g \theta_i + \varepsilon_{gi}$, where the last manipulation follows from the multiplicative relation of work pace and the baseline hazard function given in Equation (2). Note that the log-linear structure of the model guarantees the positivity of the accumulated information.

A response is given as soon as the actual level of information crosses the threshold I_g . The item-specific threshold represents the level of information that is needed in order to solve an item or to evaluate a statement from a personality scale. This

threshold is set by the knowledge demand of an item and is related to the item difficulty and the number of necessary cognitive operations. A response is given as soon as the actual level of information has reached the critical threshold. Therefore, the probability of a response before time t is $P(T_g < t | \theta_i) = P(K_{gi}(t; \theta_i, \varepsilon_{gi}) > l_g | \theta_i)$. Owing to the monotony of the logarithm this probability can be alternatively stated as $P(T_g < t | \theta_i) = P(\log[K_{gi}(t; \theta_i, \varepsilon_{gi})] > \log(l_g) | \theta_i)$. Using the relation of $\log[K_{gi}(t; \theta_i, \varepsilon_{gi})]$ to the work pace given above and simplifying further results in

$$P(T_g < t | \theta_i) = 1 - P(\varepsilon_{gi} \leq \log(l_g) - \log[\Lambda_{B_g}(t)] - \beta_g \theta_i). \quad (A1)$$

For sake of simplicity we assume that $l_g = 1$ such that $\log(l_g) = 0$ and the threshold drops out. This is not a restriction as otherwise the threshold could be absorbed into the expected amount of acquired information $\log(l_g) - \log[\Lambda_{B_g}(t)] = -\log[\Lambda_{B_g}(t)/l_g]$. In this case, the momentary rate of information acquisition is not identical to the hazard function, but only proportional. All other arguments could be applied equally.

The probability in Equation (A1) clearly depends on the distribution of the residual term ε_{gi} . For standard Gumbel-distributed residual terms ε_{gi} , the probability in Equation (A1) has the simple form

$$P(T_g < t | \theta_i) = 1 - \exp[-\exp(\beta_g \theta_i) \Lambda_{B_g}(t)]. \quad (A2)$$

The response time distribution follows from Equation (A2) by differentiating with respect to t . This gives the density implied by the proportional hazards model; see Equation (3). The identity demonstrates that the proportional hazards model has a psychological interpretation in terms of a knowledge acquisition process. The hazard function is the expected momentary rate at which individuals acquire information (or is at least proportional to it).

Appendix B

Implementation of M-Splines

The M-spline is a function that is composed of different polynomials which are defined in a piecewise manner, each polynomial for a disjoint segment of the time axis. The M-spline is generated as a linear combination of basis functions. The M-spline basis functions can be obtained as follows. Denote by $\kappa_1, \kappa_2, \dots, \kappa_K$ the interior knots that define the segment borders. In order to define the basis functions, two additionally boundary knots κ_0 and κ_{K+1} have to be chosen that border the domain of the piecewise polynomial. This could be the maximal and minimal response time. Each of the two additional knots is duplicated m times. The number of duplications determines the degree of the piecewise polynomial function. Duplicating each boundary knot m times generates a piecewise polynomial of degree $m - 1$ with $m - 2$ continuous derivatives. With the augmented knot sequence of interior knots and

duplicated boundary knots the M-spline basis functions can be generated recursively as follows. Define as the start of the recursion the functions:

$$M_{k,1}(t) = \begin{cases} \frac{1}{\kappa_{k+1} - \kappa_k}, & \text{if } \kappa_k \leq t < \kappa_{k+1} \\ 0, & \text{else} \end{cases}. \quad (\text{B1})$$

The basis functions of the M-spline basis result by using the recursion

$$M_{k,j}(t) = \frac{j \cdot [(t - \kappa_k)M_{k,j-1}(t) + (\kappa_{k+j} - t)M_{k+1,j-1}(t)]}{(j-1)(\kappa_{k+j} - \kappa_k)}, \quad (\text{B2})$$

which has to be iterated until $j = m$. The piecewise polynomial function is generated by a linear combination of the M-spline basis functions. The weights of the linear combination are restricted to be positive. This restriction guarantees the positivity of the resulting polynomial because all M-spline basis functions are positive.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Aalen, O., & Gjessing, H. (2001). Understanding the shape of the hazard rate: A process point of view. *Statistical Science*, 16, 1-14.
- Akaike, H. (1992). Information theory and an extension of the Maximum Likelihood principle. In S. Kotz & N. Johnson (Eds.), *Breakthroughs in statistics: Vol I. Foundations and basic theory* (pp. 610-624). New York, NY: Springer.
- Almansa, J., Vermunt, J., Forero, C., & Alonso, J. (2014). A factor mixture model for multivariate survival data: An application to the analysis of lifetime mental disorders. *Applied Statistics*, 63, 85-102.
- Asparouhov, T., Masyn, K., & Muthen, B. (2006, August). *Continuous time survival in latent variable models*. Presented at the Annual Meeting of the American Statistical Association, Seattle, WA.
- Baker, F., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.
- Bassili, J. (1996). The how and why of response latency measurement in telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 319-346). San Francisco, CA: Jossey-Bass.

- Bolt, D., Cohen, A., & Wollack, J. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39, 331-348.
- Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, 79, 403-425.
- Bulteel, K., Wilderjans, T., Tuerlinckx, F., & Ceulemans, E. (2013). CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods*, 45, 782-791.
- Cai, B., Lin, X., & Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics and Data Analysis*, 55, 2644-2651.
- Chernoff, H., & Lehmann, E. (1954). The use of the maximum likelihood estimates in χ^2 -tests for goodness of fit. *Annals of Mathematical Statistics*, 25, 579-586.
- Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34, 187-220.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.
- Douglas, J., Kosorok, M., & Chewing, B. (1999). A latent variable model for discrete multivariate psychometric waiting times. *Psychometrika*, 64, 69-82.
- Etezadi-Amoli, J., & Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. *Biometrics*, 43, 181-192.
- Fazio, R. (1995). Attitudes as object-evaluation associations: Determinants, consequences, and correlates of attitude accessibility. In R. Petty & J. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 247-282). Mahwah, NJ: Lawrence Erlbaum.
- Ferrando, P., & Demestre, J. (2007). Características del ítem que predicen el tiempo de respuesta en medidas de personalidad [Item characteristics that predict response latency in personality measures]. *Psicothema*, 19, 337-342.
- Fraley, C., Raftery, A., & Scrucca, L. (2014). *mclust-Normal mixture modeling for model-based clustering, classification, and density estimation* [Computer software manual, Version 4.3]. Seattle: Department of Statistics, University of Washington.
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106, 608-626.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Berlin, Germany: Springer.
- Henson, J., Reise, S., & Kim, K. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling*, 14, 202-226.
- Holden, R., & Kroner, D. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, 4, 170-173.
- Hsu, L., Santelli, J., & Hsu, J. (1989). Faking detection validity and incremental validity of response latencies to MMPI subtle and obvious items. *Journal of Personality Assessment*, 53, 278-295.

- Hughes, A., & King, M. (2003). Model selection using AIC in the presence of one-sided information. *Journal of Statistical Planning and Inference*, 115, 397-411.
- Jedidi, K., & Kohli, R. (2005). Probabilistic subset-conjunctive models for heterogeneous consumers. *Journal of Marketing Research*, 42, 483-494.
- Klein, J., & Moeschberger, M. (1997). *Survival analysis—Techniques for censored and truncated data*. Berlin, Germany: Springer.
- Klein Entink, R., Kuhn, J., Hornke, L., & Fox, J. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14, 54-75.
- Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics*, 81, 703-706.
- Li, F., Cohen, A., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurements*, 33, 353-373.
- Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in normal mixture. *Biometrika*, 88, 767-778.
- Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, 67, 304-327.
- Magidson, J., & Vermunt, J. (2001). Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, 31, 223-264.
- Mayerl, J. (2013). Response latency measurement in surveys. Detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*. Retrieved from <http://surveyinsights.org/?p=1063>
- McIntyre, H. (2011). Investigating response styles in self-report personality data via a joint structural equation mixture modeling of item responses and response times. *Personality and Individual Differences*, 50, 597-602.
- McLachlan, G., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, NY: Wiley.
- Meyer, J. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, 34, 521-538.
- Mulligan, K., Grant, J., Mockabee, S., & Monson, J. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, 11, 289-301.
- Nylund, K., Asparouhov, T., & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569.
- R Development Core Team. (2009). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: Author. Retrieved from <http://www.R-project.org>
- Ramsay, J. (1988). Monotone regression splines in action. *Psychometrika*, 3, 425-461.
- Ranger, J., & Kuhn, J. (2014). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling*, 56, 370-392.
- Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, 65, 334-349.
- Ranger, J., & Ortner, T. (2013). Response time modeling based on the proportional hazards model. *Multivariate Behavioral Research*, 48, 503-533.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

- Schneider, J., & Hübner, R. (1980). Einfluss von Verfälschungsinstruktionen auf die Bearbeitungszeit von Persönlichkeitsfragebögen [Influence of the instruction to fake on the response time in personality questionnaires]. *Zeitschrift für experimentelle und angewandte Psychologie*, 27, 565-579.
- Schnipke, D. (1999). *The influence of speededness on item-parameter estimation* (Tech. Rep. No. LSAC-R-96-07). Princeton, NJ: Law School Admission Council.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Singpurwalla, N. (2006). The hazard potential: Introduction and overview. *Journal of the American Statistical Association*, 101, 1705-1717.
- Swait, J., & Adamowicz, W. (2001). The influence of task complexity on consumer choice: A latent class model of decision strategy switching. *Journal of Consumer Research*, 28, 135-148.
- Tofighi, D., & Enders, C. (2007). Identifying the correct number of classes in growth mixture models. In G. Hancock & K. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Charlotte, NC: Information Age.
- Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, 123, 34-80.
- van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287-308.
- van der Maas, H., Molenaar, D., Maris, G., Kievit, R., & Boorsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118, 339-356.
- Varriale, R., & Vermunt, J. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, 47, 247-275.
- Vermunt, J. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17, 33-51.
- Wang, C., Chang, H., & Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66, 144-168.
- Wang, C., Fan, Z., Chang, H., & Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, 38, 381-417.
- Wenger, M., & Gibson, B. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology*, 30, 708-719.
- Wise, S., & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43, 19-38.
- Younes, N., & Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, 53, 1199-1211.
- Yung, Y. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, 62, 297-330.