**Citation: CPT Pharmacometrics Syst. Pharmacol. (2018) 7, 205–218;** doi:10.1002/psp4.12280

**TUTORIAL**

# Modeling Composite Assessment Data Using Item Response Theory

**Sebastian Ueckert**

**Composite assessments aim to combine different aspects of a disease in a single score and are utilized in a variety of therapeutic areas. The data arising from these evaluations are inherently discrete with distinct statistical properties. This tutorial presents the framework of the item response theory (IRT) for the analysis of this data type in a pharmacometric context. The article considers both conceptual (terms and assumptions) and practical questions (modeling software, data requirements, and model building).**

*CPT Pharmacometrics Syst. Pharmacol.* (2018) **7**, 205–218; doi:10.1002/psp4.12280; published online 1 March 2018.

As for many natural sciences, measurement is an essential basis for the application of the pharmacometric methodology. The comprehensive set of tools developed in this, and other, quantitative sciences becomes applicable only after physiologic or pathophysiologic quantities have been translated into numeric values. The "measurement" of disease severity is particularly challenging due to the complex, multifaceted nature of most diseases. Composite assessments strive to capture this complexity by combining different aspects of a disease in a single value. The resulting scales are essential instruments for diagnosing and monitoring of patients in many therapeutic areas. The item response theory (IRT), a statistical framework especially well-suited for the analysis of this type of data, is the subject of this tutorial.

Healthcare-related composite assessments are as diverse as the diseases they attempt to measure and especially common when no biomarkers are available. They usually take the form of a questionnaire and summarize the responses to each question in an aggregate score; however, they differ in question type, response range, assessor, aggregation method, and much more. **Table 1** gives a glance of the diversity in healthcare-related composite assessments by listing the properties of three assessments from different therapeutic areas: the Major Depression Inventory (MDI) for depression,[1] the Neuropsychiatric Inventory (NPI) for Alzheimer disease,[2] and the Expanded Disability Status Scale (EDSS) for multiple sclerosis.[3]

The origins of IRT date back to the 1940s and 1950s when standardized testing procedures, earlier mostly used by the military, were adopted for educational and achievement tests. Especially the transition from essay to multiple choice-based assessments was an important nontechnical factor for the development of IRT.[4] The theory of this approach is based on many significant advances in both psychometrics and statistics. Nevertheless, it is the works of Lazarsfeld,[5] Lord,[6] and Birnbaum[7] in the 1950s that most directly laid out its foundation. For a large-scale application of IRT, it took until the beginning of the 1980s when computational and algorithmic advances allowed the application of methodology to assessments with $>$50 questions.[4] Since then, IRT has extended to reduce the underlying assumptions further and to benefit from the widely increased computational capacity. Today, IRT is an important tool in the field of psychometrics and is used for the design and analysis of many high stakes educational testing procedures. Well-known examples include the Graduate Management Admission Test,[8] a standardized business school entry test taken annually by more than 200,000 examinees, or the Graduate Record Examination,[9] an admission requirement for most graduate schools in the United States.

When applied to healthcare-related composite assessments, IRT can provide a more powerful inference process, unique insights into the structure of the data, and enhanced simulation capabilities (we will revisit its advantages later). This tutorial aims at providing an introduction to the use of IRT to analyze healthcare-related composite assessments without choosing a particular assessment. Instead, we will focus on general concepts applicable to a broad class of outcomes. We will do so from a "pharmacometric point of view" (i.e., by taking into account the requirements of pharmacometric data analyses as well as standard tools and workflows existing in the field). At times, this will lead to a slightly different interpretation and utilization of IRT as in a traditional psychometric context. An effort will be made to highlight differences where necessary.

This article consists of two main parts: theoretical basis section and the pharmacometric application section. In the first part, we will go through the general terms, concepts, and assumptions behind IRT. In the second part, we will see how these principles apply to data commonly encountered in pharmacometrics. We will consider the choice of a modeling software, data requirements, model building, and diagnostics. In the end, we will look at more complex IRT models and examine arguments for its use in a clinical drug development setting. In addition to this main text, a supplement is provided that illustrates the implementation of a pharmacometric IRT model in nonlinear mixed-effect modeling (NONMEM).[10]

Pharmacometrics Research Group, Uppsala University, Sweden.
Received 27 June 2016; accepted 17 December 2017; published online on 1 March 2018. doi:10.1002/psp4.12280

**Table 1** Three examples of healthcare-related composite assessments that show the diversity of these classes of outcomes

| Name [Ref] | MDI | NPI | Kurtzke EDSS |
|---|---|---|---|
| **Disease** | **Depression** | **Alzheimer disease** | **Multiple sclerosis** |
| Components | 12 self-report mood questions *(e.g., "Have you felt low in spirits or sad?")* | 12 behavioral domains in dementia *(e.g., hallucinations or euphoria)* | 8 functional systems *(e.g., pyramidal or visual)* |
| Responses | 0 to 5 *("at no time" to "all the time")* | 0 to 4 for frequency and 1 to 3 for severity *("never" to "very often" and "mild" to "severe")* | 0 to 5/6 with component specific meaning *(e.g., "normal" to "quadriplegia" for pyramidal)* |
| Assessor | Patient | Caregiver | Healthcare professional |
| Aggregate score | 0 to 60 (sum of question scores) | 0 to 144 (sum of frequency times severity score from each component) | 0 to 10 in increments of 0.5 (based on a decision tree) |

EDSS, Expanded Disability Status Scale; MDI, Major Depression Inventory; NPI, Neuropsychiatric Inventory.

## Theoretical basis

We will use the example of a high school mathematical examination to explain some of the theoretical concepts and assumptions behind IRT. This use case is not only very close to its original application but also represents a simple use case that is hopefully relatable. We assume that the examination consists of several tasks each student is asked to complete and that the sum of correctly answered questions corresponds to the overall score, which is then used to rank the students.

### Basic concept

Intuitively, one will regard a student with a high examination score as being skilled in mathematics or having a high mathematical ability, despite the fact that any examination represents only a fraction of all possible mathematical problems and, more importantly, that mathematical ability is an entirely artificial construct. The interpretation of examinations or tests as surrogate measures for a hypothetical ability is, therefore, entirely part of our intuition (even if we might not always explicitly acknowledge this). One way of approaching IRT is thinking of it as a formalization of these intuitive considerations.

**Figure 1a** portrays the often implicitly made analogy between the score of the examination, let us denote it by $y_i$, and the hypothetical, and, hence, unobservable ability of the students, which we will refer to by $\psi_i$. The line, in this case, merely represents the mapping from $\psi_i$ to $y_i$, its particular shape is secondary.

When zooming in on an individual question or item, we would expect a similar principal to hold; even if the data scientist in us will want to constrain the range of the mapping to 0 (false answer) and 1 (correct answer). The latter can be achieved using a link function with the appropriate range, which, in our case, could be the frequently used expit function $(\exp(x)/(\exp(x)+1))$. On the individual item level, the stochastic nature of the process is also more tangible, and we are inclined to instead talk about "the probability of student *i* to answer item *j* correctly" or formally $P(y_{ij}=1)$. **Figure 1b** visualizes the shift from the examination, or aggregate score, to the individual item level with the adaptations discussed.

The visualization of our current concept in **Figure 1b** with the repetition of the same function for each of the three items makes one question immediately evident: Do we have to assume that the relationship between ability and the probability to answer correctly is the same across items? Of course not! It is, in fact, very likely, or even desirable, to have some questions that most students can answer correctly (item 1 in **Figure 1c**), as well as some that are intended for the best students only (item 3 in **Figure 1c**). Suddenly, items have properties that are independent of the examinees, and it is the interaction of the subject's ability and the characteristics of the item that

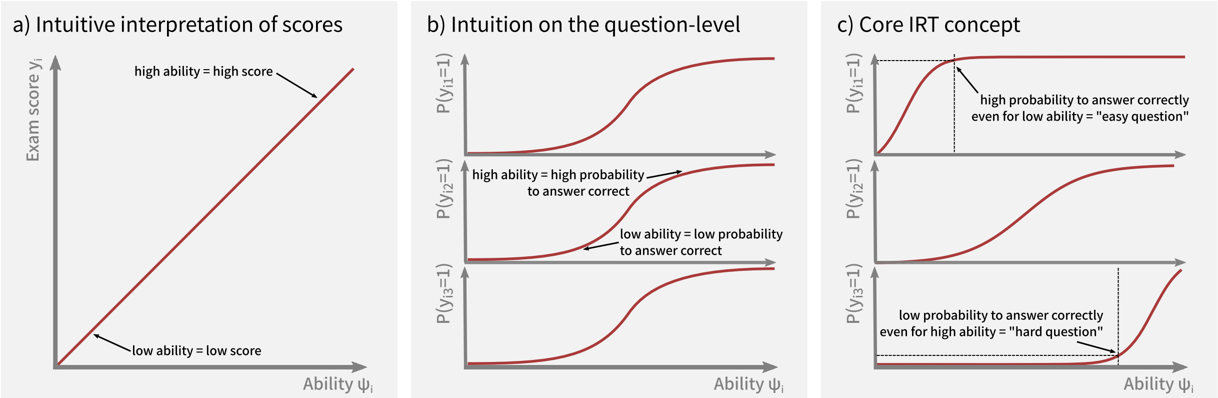
a) Intuitive interpretation of scores
Exam score $y_i$
high ability = high score
low ability = low score
Ability $\psi_i$
b) Intuition on the question-level
$P(y_{i1}=1)$
$P(y_{i2}=1)$
high ability = high probability to answer correct
low ability = low probability to answer correct
$P(y_{i3}=1)$
Ability $\psi_i$
c) Core IRT concept
$P(y_{i1}=1)$
high probability to answer correctly even for low ability = "easy question"
$P(y_{i2}=1)$
$P(y_{i3}=1)$
low probability to answer correctly even for high ability = "hard question"
Ability $\psi_i$

**Figure 1** From intuition to item response theory (IRT). The schematic illustrates how an often implied surrogacy between the examination score and ability (**a**) can be translated to the individual question level (**b**) and extended to include item-specific properties (**c**) to form the basis of the framework.

influence the response probability. This separation is the core concept of the IRT framework, and, for this tutorial, we will consider it as the defining property of an IRT model.

In fact, neither the focus on particular items nor the assumption of a hypothetical and unobserved variable is a feature that is unique to this framework. An example is a work by Hu *et al.*[11] that uses a latent variable approach to link a continuous and a discrete end point, but that one would not consider IRT modeling. What is unique about an IRT model is much more the direct treatment of characteristics of the individual assessment components (i.e., the explicit acknowledgment of the measurement process). It is this decomposition of the data into assessment-specific and subject-specific features that enable unique insights. This pragmatic framing is probably more general than IRT definitions in the psychometrics literature but it is hopefully well adapted to its pharmacometric use.

### Item characteristic functions

The nonlinear function $f_j$ that models the properties of an individual item is called an ICF, and its graphical representation is an ICC. The ICFs provide the mapping between a subject's ability $\psi_i$ and the probability of a particular response. The shape of the function is dependent on a set of item-specific parameters ($\theta_j$), and they exist not only for dichotomous responses, as in the mathematical examination example, but can be defined for ordered categorical count as well as continuous outcomes. The possibility to have varying shapes as well as response types with an assessment is the source of the enormous flexibility of IRT models to handle a variety of different assessments. Mathematically, many functions with a proper range can serve as an ICF, but a number have proven useful in practice and appear in many psychometric applications. Typically, ICFs are parameterized in a way that provides a meaningful interpretation to each of the item parameters.

For binary responses, a family of logit models with a differing number of parameters is frequently used. The simplest member of this family is the one parameter logit (1PL) model, which models the probability of a correct response from subject *i* to item *j* as:

$$P(y_{ij}=1)=f_j(\psi_i,\theta_j)=\frac{e^{\psi_i-b_j}}{1+e^{\psi_i-b_j}}=\frac{1}{1+e^{b_j-\psi_i}}, \tag{1}$$

where $\psi_i$ is the ability of subject *i* and $b_j$ is an item-specific parameter. (The model name "one parameter logit" is more evident from the mathematically equivalent formulation $\mathrm{logit}(\pi_{ij})=\psi_i-b_j$.) In the 1PL model, the parameter $b_j$ corresponds to the ability with a 50% probability of answering correctly. For a given ability, larger values of $b_j$ correspond to a lower probability to answer correctly, which is why the parameter is referred to as "item difficulty." The 1PL model is also the basis of Rasch analysis, a statistical framework closely linked to IRT but with slightly different assumptions and objectives.

The two parameter logit (2PL) model adds the item-specific parameter $a_j$ to the 1PL model and describes the probability to answer item *j* correctly as:

$$P(y_{ij}=1)=\frac{1}{1+e^{a_j(b_j-\psi_i)}}. \tag{2}$$

The extra parameter $a_j$ corresponds to the slope of the ICF at its steepest point and can be interpreted as "item discrimination" (i.e., items with a larger $a_j$ differentiate better between high and low abilities). Furthermore, $a_j$ also affects how strong two items are correlated.

The three parameter logit (3PL) model is obtained by introducing the additional parameter $c_j$ into the 2PL model, resulting in the following equation:

$$P(y_{ij}=1)=c_j+(1-c_j)\frac{1}{1+e^{a_j(b_j-\psi_i)}} \tag{3}$$

where the additional parameter $c_j$ defines a lower asymptote for the probability of a correct response. The parameter is referred to as "item guessing" as it corresponds to the probability of answering correctly even if the subject has ability 0. The effect of the three parameters $a_j$, $b_j$, and $c_j$ on the ICC is visualized in the first row of **Figure 2**.

For ordered categorical or ordinal responses with $S_j$ categories, two models are most frequently used in the IRT literature: the graded response (GR) and the generalized partial credit (GPC) model. The GR model is adapted for items that require the accomplishment of a number of tasks in which the accomplishment of one subtask requires the completion of the previous ones.[12] For our mathematical example, it could be used for a task like: "Calculate the derivative of the following equation, then use them to determine the maxima"; the second part of the task can only be performed after the first one was completed successfully. Another use-case for the GR model is graded items, for example, when the teacher scores the approach taken to solve a particular task on a scale from 0–4. The GPC model, in contrast, is adapted for tasks that consist of several subtasks that do not depend on each other.[12] It could be used for a mathematical task, such as "Solve the following 5 quadratic equations" (if we are given the subtask responses, we could also use the 3PL model for each response, or if we can assume identical ICFs for each subtask and use a binomial model with a success probability from the 3PL model).

Under the GR model, the probability for subject *i* to have at least a score of *s* for item *j* is:

$$P(y_{ij}\geq s)=\frac{1}{1+e^{a_j(b_{j,s}-\psi_i)}}, \tag{4}$$

where $a_j$ is the discrimination parameter and $b_{j,s}$ is the difficulty parameter for the s-th step of the item ($b_{j,s}\leq b_{j,s+1}$). The probability to have the score *s* is then calculated according to:

$$P(y_{ij}=s)=P(y_{ij}\geq s)-P(y_{ij}\geq s+1), \tag{5}$$

together with $P(y_{ij}\geq 0)=1$ and $P(y_{ij}\geq S_j+1)=0$. This model is most frequently used in the pharmacometrics literature to model ordered categorical data.[13]

The GPC model describes the probability to have the score *s* at item *i* as:

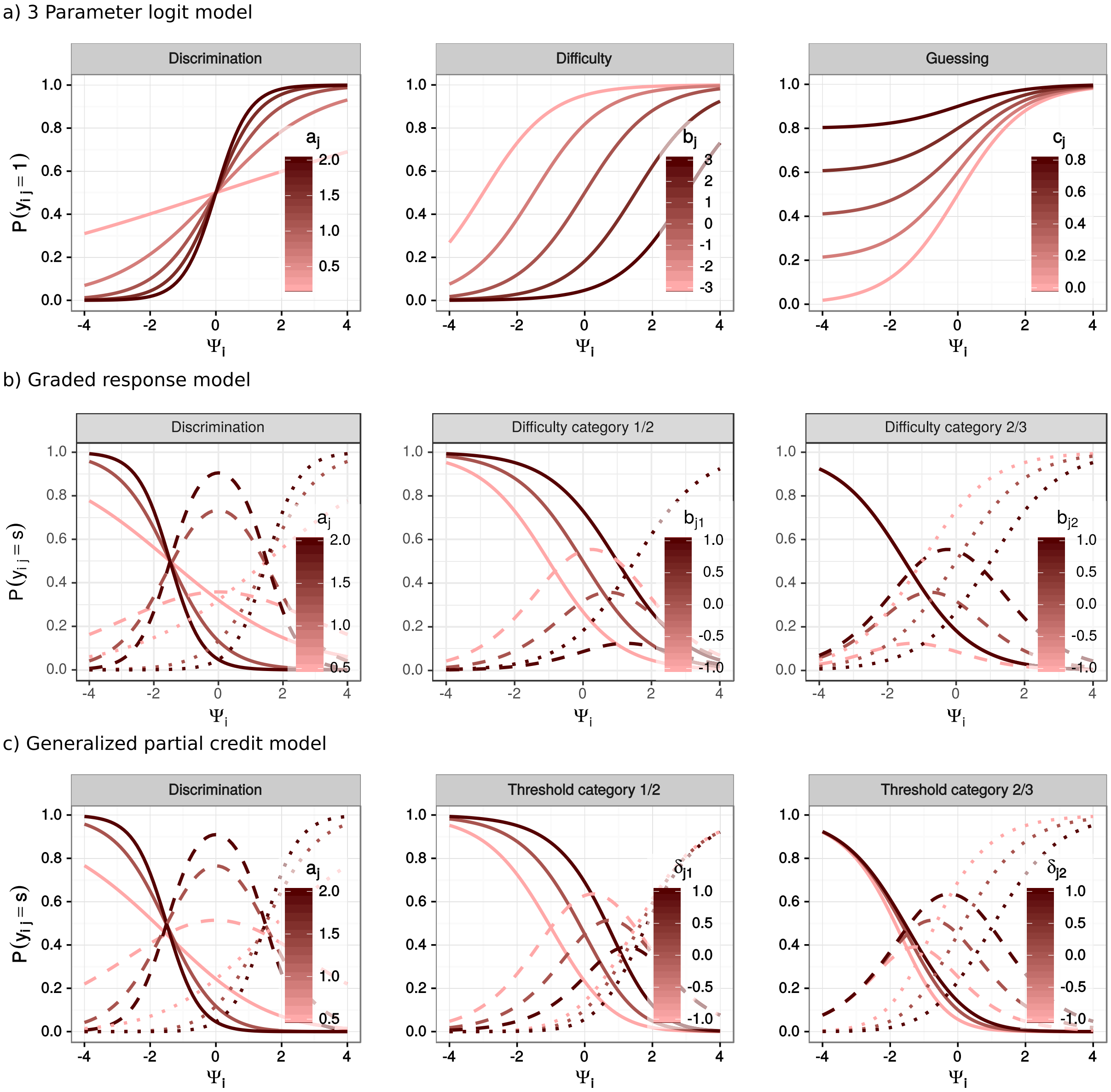

**Figure 2** Influence of item parameters on the shape of the item characteristic curves for the three parameter logit (**a**), the graded response (**b**), and the generalized partial credit models (**c**). Each panel varies one item parameter (indicated in the legend) while holding the other parameters constant. The different line types in panels **b** and **c** represent the category ($s = 1$, 2, and 3).

$$P(y_{ij}=s)=\frac{\exp\left(\sum_{l=0}^{s} a_j(\psi_i-\delta_{jl})\right)}{\sum_{m=0}^{S_j}\exp\left(\sum_{l=0}^{m} a_j(\psi_i-\delta_{jl})\right)}, \qquad (6)$$

where $a_j$ is the discrimination parameter and $\delta_{jl}$ ($l=0,\dots,S_j$ and $\delta_{j0}=0$) is the threshold parameter for the l-th category of the item.[12] The influence of the different item parameters on the shape of the ICCs of the GR and GPC model is shown in **Figure 2** rows 2 and 3, respectively.

It is not hard to think about extensions of these "classical" models and the psychometric literature provides several more complex variants for ICFs. Here, we will limit ourselves to the functions mentioned above. However, when modeling real-world data, one should not hesitate to explore different ICFs, if the ones described here do not provide a satisfactory fit.

### Item response scales

Abilities, such as in the high-school mathematical example or trait, are the most common underlying hypothetical quantities in the psychometric use of IRT. A pharmacometric application, in contrast, will utilize constructs related to the disease or health status of a subject. Here, we will use the general term "latent variable" from now on to refer to the underlying hypothetical quantity. It is maybe not immediately apparent, but the scale for the latent variable will extend from minus to plus infinity. For the mathematical example, we can understand that it will always be possible to find an even better student and that they should also have a higher value on the latent scale (similar for the opposite end of the scale). However, it should be intuitively clear that each examination has a constrained range in which we can pinpoint the location of a particular student, as soon as a student answers all questions correctly, we only get a lower bound on their ability (if we have no other information).

The ends of the latent variable scale are somehow given, but the assignment of real numbers in between is to a large degree arbitrary. For the conceptual visualization above (**Figure 1**), we could write many sequences of numbers on the scale (as long as they are linearly related) and the changes would just be reflected in different item parameters without affecting the predictions from the model. Hence, the model is unidentifiable. To make it identifiable, we need to fix the scale by defining a zero point (where does a student with an ability of zero fall) as well as a unit size (what does an improvement of one constitute). A common practice is to calibrate the scale in a reference population and define zero as the mean of that population and the unit size as its SD. In our mathematical example, therefore, a student with a score of 1 would be one SD better than the population average in the reference population. Finally, a direction for the scale has to be selected. Although this choice seems to be evident in some applications (e.g., in the mathematical example), it is less in others and needs to be carefully considered.

In summary, item response scales extend from minus to plus infinity and are relative scales, not unlike temperature scales. (If we ignore the fact that there is the absolute zero point of 0 Kelvin.) An interesting consequence is that two item response scales from different tests and populations are related through a linear transformation as long as the two tests measure the same latent variable (this requirement, however, is fundamental and in practice is not easy to establish). It is also interesting to note that 90% of the subjects of a population will be found between $-4$ and 4 if the scale has been calibrated for this population, as described above (independent of the distribution, according to Chebyshev's inequality).

### Statistical framework

In order to draw inference from real-world data using the concepts developed so far, they need to be embedded in a statistical framework, and the psychometrics literature provides multiple options for doing so. Some of these options will be more adapted for a pharmacometric use of IRT and we will restrict ourselves to the discussion of those.

Generally, we will be interested in finding predictors for the differences in reporting behavior, occasionally on the item-side, but mostly on the subject-side (i.e., we will want to explain changes in the latent variable through predictors such as time, treatment, covariates, and so on). The very flexible framework of nonlinear mixed effect models (NLMEMs) will feel familiar to most pharmacometricians and, hence, we can use pretty much any functional form to describe the relationship between predictors and the latent variable, from simple closed-form expressions to complex ordinary differential equations (ODEs).

Random effects describing between-subject variability will be included on the latent variable and, for the most part, assumed to follow a normal distribution or a transformation thereof. However, it is worth highlighting that IRT models do not depend on any normality assumption and neither does the scale calibration mentioned in the previous section. It is more the practical implementation as NLMEMs that result in a certain reliance on the normal distribution.

Finally, we can use the IRT-NLMEM either in a maximum likelihood or a Bayesian setting, with pros and cons on both sides. Here, we choose the maximum likelihood setting, but most concepts easily translate to the Bayesian formulation.

### Assumptions

The introduction already hinted at some of the advantages of an IRT-based analysis and we will revisit its benefits later. However, we need to be aware that these gains are the result of a set of assumptions implicitly made when applying an IRT model.

First of all, like for any statistical model, the inference is drawn under the assumption of a correct choice of a structural model. If, for example, the postulated ICF for one of the mathematical questions is a 2PL model (i.e., no guessing) but, in reality, there is a high chance of guessing the correct answer (e.g., multiple choice questions with only two options), then that question will bias the ability estimates toward higher values.

The second, more IRT-specific, assumption is that of unidimensionality, which postulates that the latent variable $\psi_i$ is the only shared factor between items influencing the subjects response probability. It should also be noted that it does not forbid the existence of other factors affecting the probability to a specific question. For instance, we could imagine that for the mathematical example one of the questions required some additional knowledge in physics and that students have a lower probability of answering this question correctly independent of their ability. Finally, it is essential to realize that the dimensionality of the data is dependent on both, the assessment (i.e., the items) and the subjects. A mathematical examination in which all questions demand the same level of algebra and calculus knowledge cannot distinguish between these two dimensions. However, even if the examination has varying algebra and calculus difficulty, these dimensions will also be indistinguishable if all subjects have similar algebra and calculus abilities.

The third important assumption is local independence, stating that the responses to the items given the latent variable are independent or that a response of one item does not change the probability for a response of another item (conditional independence). In practice, several possible factors can introduce an additional dependence between items. For example, a student's confidence might change if he receives immediate feedback about his answers (like in an oral examination). The assumption of local dependence is conceptually similar to unidimensionality but affects pairs of items instead of the whole assessment.

**Table 2** Items of the fictitious RA score used as an illustration example and their response range

| | Item | Range |
|---|---|---|
| 1 | Patients global assessment of disease activity | 0–5 |
| 2 | Pain | 0–5 |
| 3 | "Dress yourself" | 0 (able)/1 (not able) |
| 4 | "Get in and out of bed" | 0 (able)/1 (not able) |
| 5 | "Walk outdoors on flat ground" | 0 (able)/1 (not able) |
| 6 | "Get a good night sleep" | 0 (able)/1 (not able) |
| 7 | "Turn regular faucets on and off" | 0 (able)/1 (not able) |
| **Total score** | | 0–15 |

As with any modeling effort, some or all of these assumptions will be violated in practice (to a certain degree). However, it is crucial to strive to minimize these violations to reduce their impact. We will discuss some diagnostics that allow evaluating assumptions in section model diagnostics, and in section advanced topics and extensions we will highlight a few extensions to the IRT concept, as introduced here, that relax the assumptions mentioned.

## PHARMACOMETRIC APPLICATION

In the second part of this tutorial, we will use the framework outlined so far and apply it to pharmacometric data. Like before, we will use a small example to illustrate the individual steps necessary to build the model as well as to highlight potential problems. For this illustration example, we will use a hypothetical assessment with seven items (2 rating scale items and 5 yes/no questions) as it might be used to assess the severity of rheumatoid arthritis (RA). The individual items of this hypothetical assessment are summarized in **Table 2**.

As mentioned before, we will focus on general concepts rather than describing the implementation of a particular example. However, the supplementary material of this tutorial contains some guidance on how to implement a simple IRT model for the RA example in NONMEM.

### Software

Several commercial and noncommercial computer programs specialized on the development and use of IRT models exist. The advantages of using software specifically designed for IRT-modeling include a more straightforward implementation of the model, especially adapted estimation algorithms and automatic generation of standard diagnostics. On the other hand, these programs were developed for a psychometric application and, from a pharmacometric point of view, might lack flexibility (it can be challenging or impossible to include ODEs) as well as require additional time to learn the use of the program. An overview of the available programs can be found on the Wikipedia page for "psychometric software" (en.wikipedia.org/wiki/Psychometric_software).

There are also several R packages for the implementation of IRT models available. These packages have the general advantage that installation and, hence, testing are simple and that they operate in an environment most pharmacometricians are already familiar with. The CRAN task view for "Psychometric Models and Methods" is a good way to explore available packages. Especially recommended from this list is the package "MIRT,"[14] which provides an impressive set of functionality and to develop models even for large assessments quickly. However, similar to the specialized stand-alone IRT software, these R packages (including "MIRT") are very focused on psychometrics and lack flexibility for modeling longitudinal data.

General purpose NLMEM modeling software, such as NONMEM,[10] SAS (SAS Institute, Cary, NC), or Stan[15] (the Stan user manual[16] contains an example of how to implement IRT models) are an alternative for the development of IRT models. They do not provide the same level of assistance as specialized software regarding coding the ICFs or generation of diagnostics but allow for a much higher level of flexibility.

### Data and data format

A large number of parameters in the model and a low amount of information per item result in the requirement of a high number of observations for the precise estimation of item parameters. The exact sample size will depend on a complex interplay between the number of items in the assessment, their information content, the model structure, the heterogeneity of the subjects etc. DeMars[17] gives some general guidance on the number of subjects required. For a 2PL or 3PL model with fixed $c_j$ parameter, at least 500 subjects are needed, 1,000 subjects are considered a more cautious guideline if more complex models or distributions are to be estimated and although more subjects will lead to higher precision, little benefit is found beyond 2,000 subjects. For ordered categorical items, the number of subjects per category is a critical determinant for parameter precision (ideally all categories occur as responses). DeMars reports that samples as small as 250 subjects were sufficient for a 3 category model but that this number increased to more than 1,000 when parameters for 6 categories were to be estimated. It should be noted, however, that these guidelines are valid for cross-sectional studies and that longitudinal data generally provides more information per subject.[17]

The response data needs to be available with an item-level resolution for all subjects and visits to develop an NLMEM IRT model. The chosen modeling software will largely determine the particular formatting of the dataset. Missing responses at a specific visit (i.e., a patient did not finish answering the questionnaire), can be ignored when they are missing completely at random. However, it is advisable to include them in the dataset to be able to test different strategies for handling missing data. On the longitudinal level, missing data for IRT models does not differ from other pharmacometric models and the same care regarding the missing data mechanism needs to be taken.

### Analysis steps and model building

Our goal in the model building process will be to describe the probability of each subject's response over time and under treatment. Hence, we want to build a model of the form as follows:

$$P(y_{ijk}=s)=f_j(\psi_i(t_k,\cdot),\theta_j) \qquad (7)$$

where $y_{ijk}$ is now the response from subject $i$ to item $j$ at visit $k$ and $t_k$ is the time of that visit. The notation $\psi_i(t_k,\cdot)$ is used as it highlights that it will be our goal to explain changes over time primarily through changes in the latent variable and consider the ICFs $f_j$ static.

Similar to the stepwise development of the pharmacokinetic and the pharmacodynamic component of a pharmacokinetic/pharmacodynamic model, it can be helpful to break down the development of a complex IRT model into the following two steps: in step one is the item response component and in step two is the longitudinal component.

Step one focuses on the choice of a model describing the relationship between latent variable and item responses (i.e., $f_j$ and $\theta_j$ in Eq. 7), while keeping the structure of the latent variable model $\psi_i(t_k,\cdot)$ fixed. The modeling of this component mainly consists in the choice of appropriate ICFs and we will discuss the particularities of the model structure in the Item Response Component section. During the development of the item response component, the latent variable model $\psi(t_k,\cdot)$ acts only as a placeholder and will later be replaced with the longitudinal model. Nevertheless, the choice of the structure for the placeholder can have a profound influence on the outcome of the analysis, and it should, therefore, be taken with care. The pros and cons of some possible unstructured longitudinal models will be discussed in the Unstructured Longitudinal Model section.

The longitudinal component developed in step two describes the change of the latent variable over time or under treatment (i.e., a model for $\psi_i(t_k,\cdot)$ will be chosen while keeping the ICFs $f_j$ and their item parameters $\theta_j$ fixed to structure and estimates from step one). We will discuss this modeling step in the Longitudinal Model section and the Exposure-Response Model section. At the end of the two-step procedure, a re-estimation of all parameters in the joint-model as well as re-evaluation of important diagnostics has to be performed.

Alternatively, it is also possible to develop the longitudinal and the item response component simultaneously (i.e., to make adjustments to both item response component and the longitudinal component at the same time). Although this approach has some indisputable advantages, such as the best fit, it can be challenging to attribute model misspecifications correctly, and model stability might be an issue. Therefore, the simultaneous model development is probably best suited if data availability is expected to limit longitudinal model complexity.

No matter if we choose a simultaneous or two-step model development, the model building process will be similar to other pharmacometric models: incremental model modifications are performed on a simple base model and subsequently evaluated using a diverse set of diagnostics (described in the Model Diagnostics section). This cycle of improvement ends when the model performs satisfactorily for its intended purpose.

### Item response component

The item response model captures the relationship between item-level observations and latent variable. In consequence, it also defines the actual IRT scale (Item Response Scale section) and is of central importance for all further inference. It is not always necessary to develop the item response component model from scratch, sometimes an appropriate model for the assessment at hand might be available in the literature (e.g., as the work of Balsis *et al.*[18] for the ADAS-cog scale). In other instances, the assessment was even developed using IRT and the results from that analysis can be used. Most of the time, however, the same dataset will be used to develop both item response and longitudinal components. In these cases, it is important to recall the relative nature of IRT scales (as described in the Item Response Scales section). The subjects available in the data become the reference population, which not only has consequences for the inference but also affects the model building, as we will discuss later.

The choice of the ICFs $f_j$ for each item might seem to be a formality (i.e., as entirely defined through the data type). In most cases, however, there are several plausible choices of ICFs per item, each of them with a slightly different set of assumptions. For example, we could either model items three to seven of the RA score as binary responses or describe the sum of “not able” responses using an ordered categorical model. Although the former provides more information, it also includes an additional independence assumption (conditional on the latent variable). Even if we have decided to treat the items as binary, we still have to choose among the 1PL, 2PL, or 3PL models. For our RA score in **Table 2**, some of the binary items do not appear to be very disease-specific (such as the quality of sleep). Therefore, we might directly assume a non-zero probability of a “one”-response even for subjects with low latent variable values and use a 3PL binary model. On the other hand, if the data source is rather small or mostly contains subjects with a high disease severity, we might instead start with a 2PL model and evaluate if a non-zero $c_j$ parameter significantly improves the model fit.

After the ICFs have been chosen, we can explore the effect of covariates on the item parameters. For the RA example, maybe we would discover that patients who live in an assisted living environment tend to report less problems for environment-specific items (“Get in and out of bed” and “Turn regular faucets on and off”) or that the patients from Japan report lower pain scores than their European counterparts.

In some cases, we might even want to evaluate a time-dependent relationship between latent variable and response through time-varying covariates; for example, to capture learning effects in a cognitive assessment.

The points above illustrate the potential complications when developing the item response model and they highlight that knowledge about the structure of the assessment, the nature of the items, and the limitations of the available data are indispensable during model development.

### Unstructured longitudinal model

If the two-step model development strategy is selected, an unstructured longitudinal model can serve as a placeholder during the development of the item response model. The simplest choice is to select one visit (e.g., baseline) from the available data and use it to build the item response component (i.e., to use a cross-sectional slice of the data). With that choice, the latent variable model is merely as follows:

$$\psi_i(t_1)=\eta_i \tag{8}$$

where $\psi_i(t_1)$ denotes the latent variable at the first occasion and $\eta_i$ is a normally distributed random variable with mean 0 and variance 1 (i.e., $\eta_i \sim N(0,1)$). The fixed mean and variance might be surprising at first; however, it corresponds to the mathematical implementation of the scale definition described in the Item Response Scales section. Essentially, we define the subjects in the data at the chosen visit as our reference population and set zero point and unit size for the scale accordingly. (Technically, there is also an identifiability issue for most ICFs among latent variable, item difficulty $b_j$, and item discrimination $a_j$ i.e., it is impossible to estimate $x$ and $y$ if they appear only as $x$ and $y$ in a model.) This strategy is simple to implement and does not require the definition of a longitudinal placeholder model. Hence, there is no risk of misspecifications from the longitudinal model to affect the item parameters. The main disadvantage of this strategy is the omission of the majority of the data, leading to a higher uncertainty in the ICF parameters. Furthermore, the baseline visit is often much more homogeneous than later visits and might not cover the full range of latent variable variability. The strategy is, therefore, best suited when a large (relative to the complexity of the model) external data source is available for the development of the item response component. It is noteworthy that the decomposition into item and subject-specific characteristics, the core concept of IRT, allows us to use a much wider range of data sources than maybe with other approaches. For example, it is possible to combine populations with different baseline characteristics, disease levels, and background drug therapies as long as it is reasonable to assume that the ICFs are the same across populations. We can use some of the diagnostics described in the next section to verify this critical assumption. An ideal external data source would cover a broader population than the actual analysis dataset, this way we can be more confident when extrapolating to more or less severe patient populations and we will get higher precision in item parameter estimates.

A simple strategy of using all the data when building the IRT model is to treat each visit independently, resulting in the latent variable model as follows:

$$\psi_i(t_k)=\eta_{ik} \tag{9}$$

with

$$\eta_{ik} \sim \begin{cases} N(0,1) & \text{if k=1} \\ N(\mu_k, \omega_k^2) & \text{otherwise} \end{cases} \tag{10}$$

where $\psi_i(t_k)$ is now the latent variable model for the visit at time $t_k$, and $\mu_k$ and $\omega_k^2$ are mean and variance at that visit. Although this strategy utilizes all the data, it might still be not suitable for uninformative assessments and is complicated to implement if occasions are not clearly defined (i.e., times of the visits differ significantly between subjects).

### Longitudinal model

The longitudinal component describes the evolution of the latent variable w.r.t. time or treatment and its complexity can range from simple linear models to complex nonlinear functions described via ODEs. Unusual for modeling might be the range from minus to plus infinity, which, depending on the chosen model, might need to be taken care of. For example, a simple linear model for the RA example, such as:

$$\psi_i(t)=\text{base}_i+\text{slp}_i \cdot t, \tag{11}$$

would not require any modification (i.e., the function already covers the right range). However, if we were to choose a more complex, semimechanistic, indirect-response type model[19] of the form:

$$\frac{dR}{dt}=k_{in_i}-k_{out_i}\cdot I(t)\cdot R \quad R(0)=\frac{k_{in_i}}{k_{out_i}}, \tag{12}$$

with the intention to replace $R$ with $\psi_i$, then we need to handle the fact that $\psi_i$ can be negative. A possible implementation could be:

$$\frac{d\psi_i}{dt}=k_{in_i}-k_{out_i}\cdot I(t)\cdot\left(\psi_i+\frac{k_{in_i}}{k_{out_i}}-\psi_{0_i}\right) \quad \psi_i(0)=\psi_{0_i}, \tag{13}$$

where $\psi_{0_i}$ is now allowed to be smaller than zero.

The uncommon range together with the nonlinear transformation through the ICFs can make it also challenging to infer an appropriate longitudinal model (see **Figure 3**). It can, therefore, be helpful to plot empirical Bayes estimates (EBEs) of the latent variable, from a model similar to the one described in Eq. 9, vs. time, to obtain an understanding of the dynamics on the latent variable scale.

Another potential challenge is respecting the choice of the zero point and the unit size for the IRT scale during longitudinal model development (Item Response Scales section). In most cases, this comes down to the specification of the distribution for the latent variable at baseline (e.g., base$_i$ and $\psi_{0_i}$ above). If, for example, the analysis population at baseline were defined as having a mean of 0 and variance of 1, then we need to keep those assumptions when estimating the longitudinal parameters ($\text{base}_i, \psi_{0_i} \sim N(0,1)$), or the model will be unidentifiable during a subsequent joint estimation of item and longitudinal components. It is worth noting that these complications disappear if an external data source was used to develop the item response component, as described in the Item Response Component section. In the case of the linear model for the RA example, the baseline parameter base$_i$ would merely be relative to the population from the external source (i.e., a mean $-1$ and a variance of 0.5 would tell us that the analysis population is less severe and less diverse than the external reference).

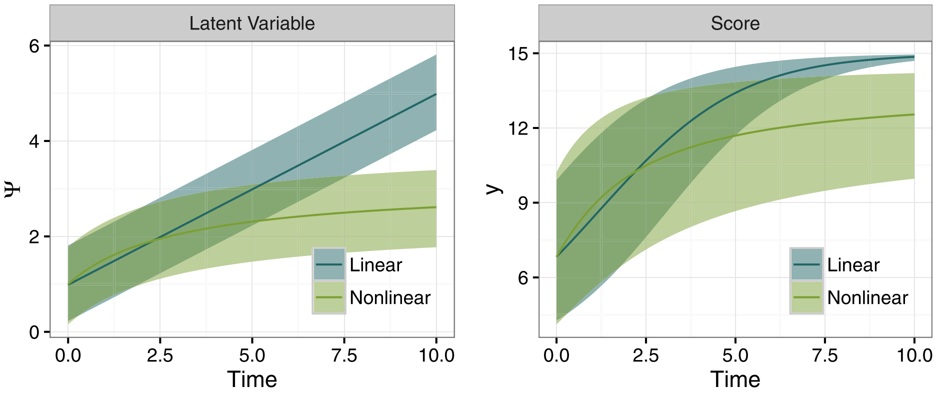

**Figure 3** Simulations (mean and 95% prediction interval) from a linear and nonlinear longitudinal model on the latent variable scale and the resulting longitudinal evolution of the score for the rheumatoid arthritis score example. The transformation through the item characteristic functions make inferring the underlying longitudinal model from the score alone challenging.

Both time-constant and time-varying covariates can easily be included on the longitudinal model component and help to explain why some subjects progress faster or start with a higher value at baseline. Again, care needs to be taken that eventual scale definitions are kept. For example, if the baseline value for the linear RA model from above is supposed to differ between women and men, but the item parameters have been fixed under the assumption of mean 0, variance 1 for the joint population, then one needs to express mean and variance of the latent variable for one sex in function of mean and variance of the other. Already, for this simple, bivariate case, the equations for mean and variance of males ($\mu_m, \sigma_m^2$) as a function of the values for females ($\mu_f, \sigma_f^2$) and the fraction of females in the population ($\pi_f$) is reasonably complex (These equations assume mean 0, variance 1 in the joint population. They can be obtained by solving the equation for the mean and variance of a mixture of distributions for one of the mixtures.):

$$\mu_m = -\frac{\pi_f}{1-\pi_f} \cdot \mu_f \tag{14}$$

$$\sigma_m^2 = \frac{1-\pi_f(\mu_f^2+\sigma_f^2)}{1-\pi_f} - \mu_m^2 \tag{15}$$

We might, therefore, instead define one particular covariate value as the reference (i.e., the one for which mean 0, variance 1 holds), and estimate the other means and variances relative to it (the ICFs parameters need to be re-estimated when doing this switch).

A noteworthy technique that avoids many of the scale normalization issues is to fix the difficulty and discrimination parameters of one of the items instead of the mean and variance of the latent variable. This way, the identifiability issue is removed and it is possible to estimate mean and variance as usual.

## Exposure-response model

Investigating the link between exposure and response is a key consideration for most pharmacometric modeling exercises. In an IRT model, it seems most natural to do this on the latent variable (i.e., to change the disease state or the progression of the disease while leaving the properties of the assessment unchanged). As for the longitudinal model, the range of the latent variable, as well as the distortion through the ICF, might need some time to get used to. However, the usual questions in regard to exposure-response modeling still apply (i.e., also in an IRT-based analysis the modeler needs to investigate: "Which exposure metric is the best predictor for the response?" "What functional form describes the relationship best?" "Is there a delay between exposure and response?" etc.). With the two example disease progression models for the RA example from the previous section in mind, we could imagine to investigate a linear dose-response relationship for the slope parameter slp$_i$ in the linear model or to test an maximum effect ($E_{max}$) concentration effect on the catabolic rate parameter $k_{out_i}$ in the indirect-response model.

Even if the latent variable seems to be the most obvious place to investigate an exposure-response relationship, it is still possible to test for an additional drug effect on individual or groups of items. This way, we could, for the RA example, test whether a treatment has a stronger effect on pain than explained through changes in the latent variable alone. However, it should be noted that a difference in drug effect for different items might be an indicator for a violation of the unidimensionality assumption (Assumptions section).

## Model diagnostics

The IRT model diagnostics need to evaluate the different model components and verify the underlying assumptions. There is a wealth of numerical and graphical diagnostics available and it is important to acknowledge that each highlights a particular facet of the model best. With this in mind, we can group the available diagnostics as follows: (i) diagnostics for the ICF fit to a particular item; (ii) diagnostics for pairs of items; and (iii) diagnostics for the overall goodness of fit (GOF). For each of those groups, the following sections provide an overview of the most important diagnostics.

In addition to the descriptions given in the following section, the graphical diagnostics applied to the RA example are shown in **Figures 4, 5, and 6**. The figures focus on a few items and the remaining ones can be found in the electronic **Supplementary Material**. All diagnostics displayed

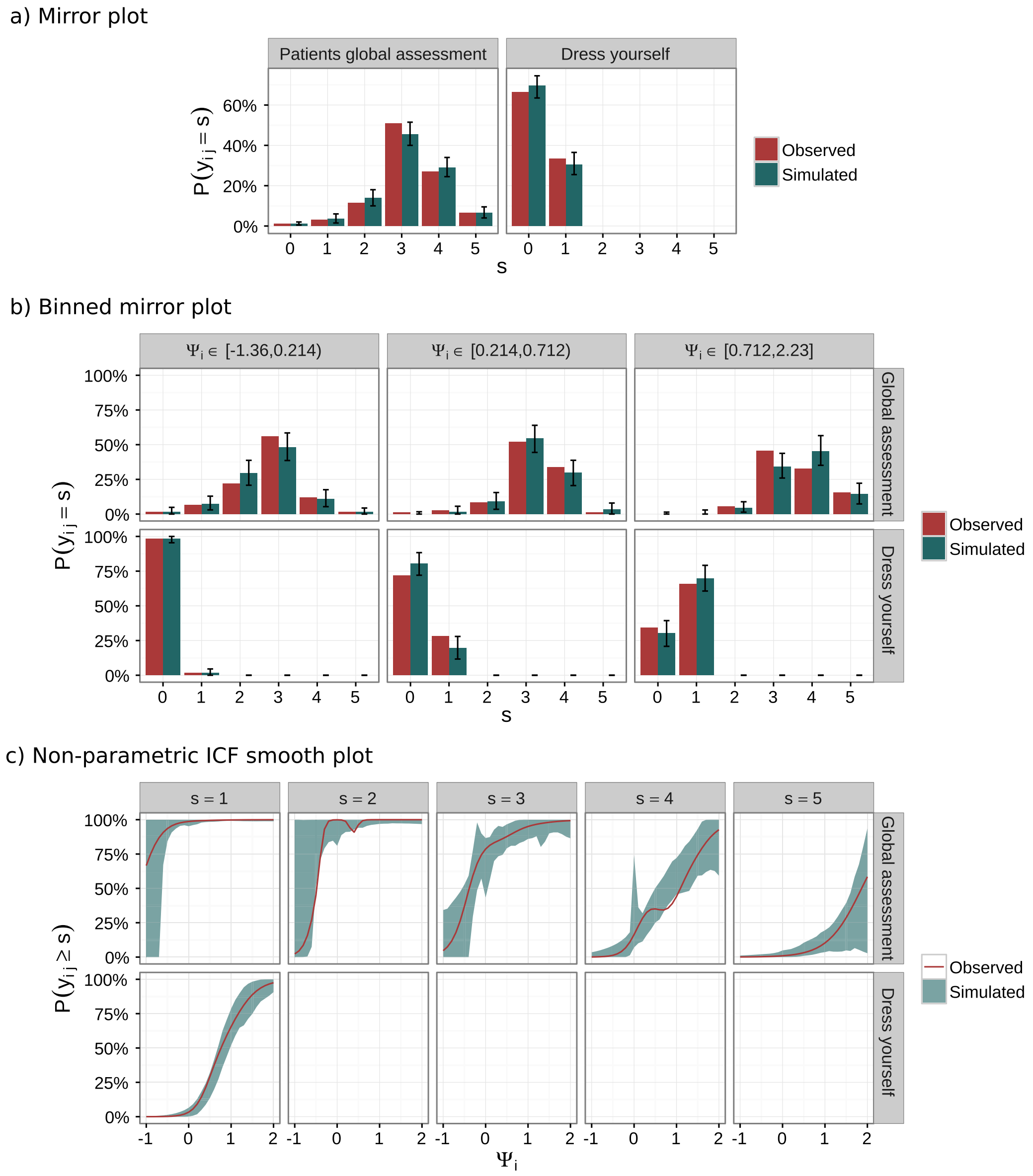

**Figure 4** Mirror plot (**a**), binned mirror plot (**b**), and nonparametric item characteristic function smooth plot (**c**) as examples for graphical item-level diagnostics. All diagnostics are shown for two items of the rheumatoid arthritis example and are based on the “true” analysis model.

have been generated for a simulated data with the “true” analysis model (i.e., the one that has been used to simulate the data) and, hence, are not expected to display any model misspecification.

### Item-fit diagnostics

Item-fit diagnostics focus on diagnosing the relationship between latent variable and response while ignoring the dynamics of the latent variable. They can, therefore, be interpreted as verifying the assumption of correct ICF fit. Depending on the chosen modeling strategy and the development stage of the model, these diagnostics need to be generated at different time points to verify the time-invariance of the item response component.

A simple graphical diagnostic for the GOF of an individual ICC can be generated by plotting the distribution of responses at a particular visit together with the distribution of one or multiple simulations from the model (“mirror-plot”). The multiple simulations version for two of the RA items at baseline is shown in **Figure 4a**. Although this diagnostic is easy to communicate even to nonmodelers, it only diagnoses the description averaged over the population and, therefore, has a rather low power to detect misspecifications. This is especially apparent for binary data in which the diagnostic shows the fraction of zero or one response (i.e., a single value per item), whereas there are two (2PL), three (3PL), or even more parameters estimated for each item.

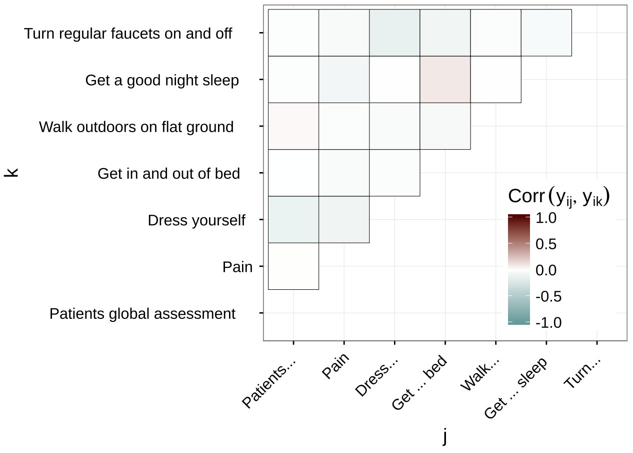

**Figure 5** Residual correlation plot between all pairs of items of the rheumatoid arthritis example based on the “true” analysis model.

A more informative diagnostic can be obtained by stratifying this plot by the EBEs for the latent variable, which can be obtained both for the observed and simulated data (with an additional estimation step for the latter). A version with three strata for the same items of the RA data example is shown in **Figure 4b**. In contrast to the simple mirror plot, the diagnostics now allow identifying which part of the ICF is potentially misspecified. In practice, the choice for the number of bins needs to be adapted to the number of subjects in the dataset as well as their heterogeneity; however, a poor choice might hide a misspecification.

An extension of this graph that avoids binning is the nonparametric ICF smooth plot. It uses nonparametric smoothing splines to fit responses and EBEs for both observed data and multiple simulated datasets. The data-based smoothing spline is then plotted together with the 95% confidence band of the simulation-based smooths and expected to fall within this band for an acceptable model fit. Generalized additive models (GAMs) using the binomial distribution are particularly suited for this purpose when dealing with binary or ordered categorical data. **Figure 4c** shows a GAM-based version of this nonparametric ICC smooth plot for the RA data. In this figure, the smoothing parameter for the GAM fit is obtained through cross-validation. Although conceptually rather complex, these graphs are easily producible for example using the “GAM” R package.[20] This diagnostic has the advantage to visualize the ICCs of the items and, therefore, allow to direct the model building process better.

It should be noted that both mirror and GAM-smoothed diagnostics described in this section utilize the EBEs of the

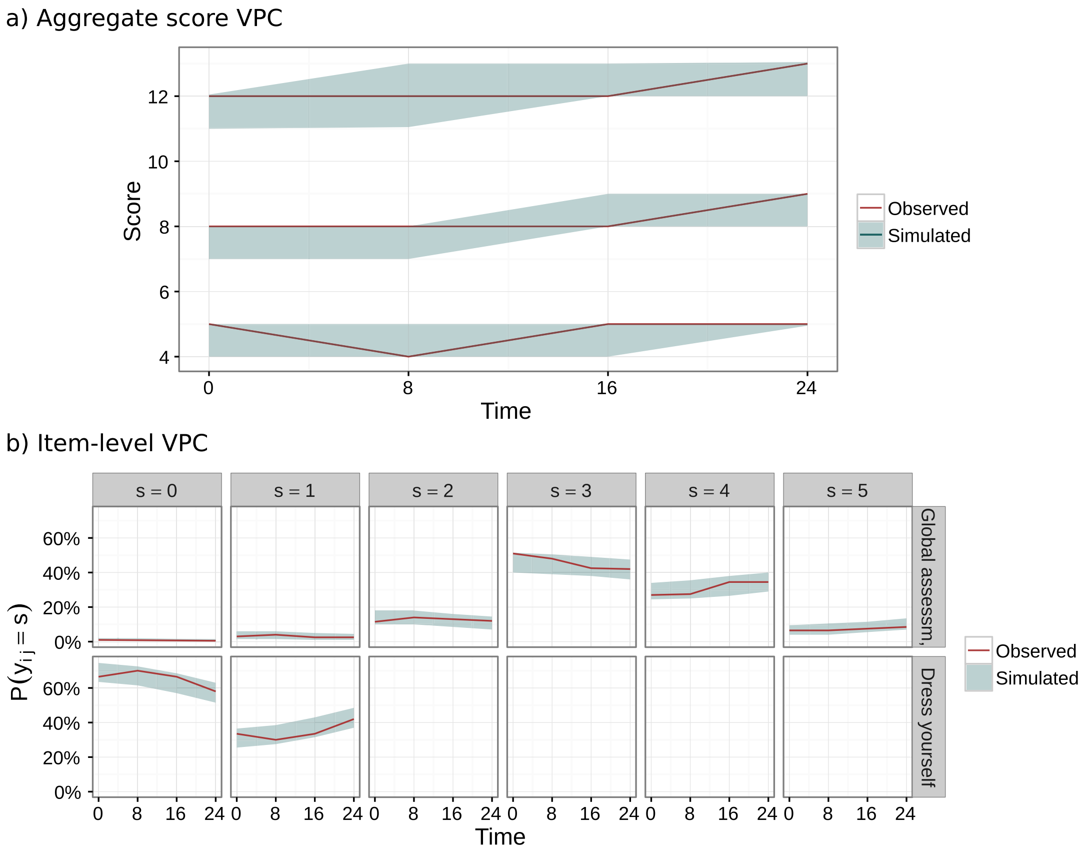

**Figure 6** Aggregate score visual predictive check (VPC) for the rheumatoid arthritis score showing the median, 2.5th, and 97.5th percentile of the observations together with the corresponding confidence intervals from the model (**a**) and item-level VPC for two of the rheumatoid arthritis score items (**b**), both based on the “true” analysis model.

latent variable for the observed and simulated data, hence, there are expected to be less affected by shrinkage. Nevertheless, it is advisable to check shrinkage values for the random effects associated with the latent variable when interpreting these diagnostics.

A numerical diagnostics that evaluate the fit of individual items is the S-$\chi^2$ statistic,[21,22] which tests the difference between observed and expected (i.e., model-based) item responses. A significant difference between the two is an indicator for a misspecification. As the S-$\chi^2$ statistic is to be calculated for each item, a family-wise error rate correction (such as the Bonferroni correction) has to be performed.

After misspecifications for a particular item has been identified, we can choose a different ICF to try to correct it. Finally, if a satisfactory fit cannot be obtained, it is also an option to remove the item from the analysis. However, this rather drastic step should only be considered if removal can also be justified.

### Item-pairs fit

Diagnostics for the fit of item pairs are of particular importance as they can be used to diagnose the misfit of a specific item (if an item appears in several misspecified pairs, it is likely to be the culprit) as well as to test the assumptions of conditional independence and unidimensionality.

Graphically, the fit of pairs of items can be evaluated using residual correlation plots.[23] The residuals are obtained as the difference between expected and observed responses for each subject and item, and they may be standardized by dividing by the expected SD. We can then visualize the correlation matrix of the residuals between pairs of items. For a correctly specified model, we would expect no significant correlation between residuals. In **Figure 5**, the residual correlation plot for the RA score at baseline is plotted. The residual correlation plot can also be used to identify longitudinal correlations between pairs of items that are not explained by the model.

Numerical diagnostics for pairs of items are popular in the psychometrics literature and exist in many different versions. These statistics are generally all based on residuals but differ in the way the residuals are calculated and weighted. An overview of the available statistics gives the work by Maydeu-Olivares,[24] which found that for binary and ordinal data the *z* statistic has the best type I error and power behavior. When evaluating the significance of the misspecification of all possible pairs of items, a family-wise error rate correction, which, again, has to be taken into account.

Misspecifications identified through item-pairs-fit diagnostics can either be resolved by modifying the ICF for some of the affected items or might be an indicator that more latent variable dimensions are required. We will briefly discuss the extension of IRT models to multiple latent variables in Advanced Topics and Extensions section.

### Overall GOF

Overall GOF diagnostics take both the fit of the items and the fit of longitudinal model into account.

The most versatile graphical diagnostic for the general model fit is the visual predictive check (VPC). Like for other pharmacometric models, VPCs can be generated as a function of time or other predictors and should be stratified by important covariates. A particularity for IRT models is that VPCs can be generated both on the item-level and on the aggregate score scale. Generally, the former will show the evolution of different response probabilities over time and the later the evolution of certain percentiles (often median, 2.5th and 97.5th percentile) of the total score. Response probability and percentiles for the observed data are then shown together with the corresponding confidence intervals (often the 95% confidence interval) from the model. It is important to evaluate the model on both scales, because a good fit for the summary score can be the result of positive and negative biases for different items that cancel each other out. **Figure 6** shows aggregate score (**Figure 6a**) and item-level based (**Figure 6b**) VPCs for the RA example.

There are two classes of numerical diagnostics to evaluate the overall GOF of an IRT model. The first class of diagnostics is test-based (i.e., they test the hypothesis that the data is consistent with the distribution specified by the model). For IRT models, the classical Pearson's $\chi^2$ test statistic is not appropriate and limited information GOF statistics are commonly used.[25] The second class of diagnostics evaluate the closeness of the developed model to the unknown truth. A popular statistic in the psychometric literature is the standardized root mean square residual, which allows the definition of closeness of fit criteria independent of the model complexity.[26] At this point, it shall also be mentioned that the GOF between two competing nested models can be tested using the commonly used likelihood ratio test.

Overall, model fit misspecifications can result from any of the model component and it can be difficult at times to pinpoint its origin. It is, therefore, essential to also evaluate the diagnostics described in the previous sections.

### Advanced topics and extensions

In this tutorial, we only scratched the surface of the large class of IRT models and focused on rather basic models. These models often come with assumptions that are hard to justify in practice. Fortunately, there are a number of extensions that allow the application of IRT also in more complex situations.

The assumed unidimensionality is often an issue. Scores might consist of multiple subcomponents that are intended to measure different aspects of a disease. The Positive and Negative Syndrome Scale (PANSS) score in schizophrenia, for example, consists of positive, negative, and general PANSS subscales with potentially different evolution and sensitivity to drug effects. In their IRT model, Krekels *et al.*[27] used a separate latent variable for each of the subcomponents to take this into account. Rather than using an *a priori* grouping of items, Gottipati *et al.*[23] used IRT-based residual diagnostics to identify three (one of which with a mixture) latent variables for the Movement Disorder Society-Sponsored Revision of the Unified Parkinson Disease Rating Scale (MDS-UPDRS) scale in Parkinson disease. In some situations, the existence of separate components might not even be tangible. An example is the

NPI[2] (see **Table 1**). In an earlier work, we demonstrated the applicability of multidimensional item response theory (MIRT) for these types of composite assessments.[28] In short, MIRT lets the response for one item depend on a function of several latent component while allowing a different contribution of components for different items.[29]

The assumption of conditional independence between items is another simplification made throughout this tutorial that might be violated in practice. For the RA score example, we could imagine a dependence between "pain" and "walk outdoors on flat ground" at a given time point or between the ability to "get in and out of bed" at two time points, that is not explained by the latent variable. The introduction of additional random effects that are shared between groups of items is one possibility to handle this conditional dependence. The resulting type of models are referred to as "testlet" models in the field of psychometrics (the dependent items form a testlet).[30] In addition to a group of items at the same visit, conditional dependence might also occur for the same item over time, especially when observations are frequent. Germovsek *et al.*[31] recently demonstrated how to handle longitudinal dependence for a patient-reported outcome score reported on a daily basis by linking an IRT model to a continuous-time Markov model.

For completeness, it shall also be mentioned that there is a whole class of IRT models that relax the assumption of a correct model fit or more precisely that do not assume a particular parametric model. An introduction to nonparametric IRT can be found, for example, in the book by Sijtsma & Molenaar.[32]

### When to use a pharmacometric IRT model

An IRT model for the very short RA score example requires probably around 30 parameters, an NLMEM for the sum of that score, on the other hand, could be developed with as few as four parameters. This difference in the number of parameters hints at the increased complexity of an IRT-based analysis and might substantiate the question: Is this increased effort worth it? As often, the answer to this question is: it depends. The IRT models indeed provide unique insights, but not all pharmacometric problems necessitate those insights. We will, therefore, use this final section to highlight some pharmacometric problems that could particularly benefit from an application of IRT, but we will also discuss some drawbacks of this approach.

The IRT models allow us to separate the properties of an assessment from the characteristics of the patient population and the influence of the disease. Hence, they provide a natural framework to combine different outcomes from the same disease into a joint disease model. We can more easily pool data from studies with different assessment variants, treat different endpoints in a study as observations of the same underlying disease variable, or link trial-specific outcomes to clinical-routine tests.[33,34] We also have greater liberty to adapt the assessment to the patient population in future trials or even modify the assessment dynamically as the patients progress, and still maintain a common reference.

The knowledge about the assessment in itself can be very valuable. Rather than being a mere black-box the structure and administration of an assessment become part of the trial design space that can be optimized. We could, for example, select the most sensitive items for a particular disease severity to streamline the assessment process,[35–37] or eliminate assessment components that have a disadvantageous signal to noise profile.[38] Our understanding of an assessment can be even further increased by identifying covariates. We might then be able, for example, to distinguish whether it is disease severity or reporting behavior that differs between patients from different geographical regions.

With items as the elemental unit, our models also gain in flexibility. Hence, we might be able to identify components of the disease with differing natural progression or could test if a treatment effect differs for individual or groups of items.[27,38,39]

Last but not least, once we have established an IRT model for an assessment, we can use this knowledge for our benefit even when analyzing new data. For example, we could only re-estimate the parameters that describe the dynamics of the latent variable from the new data while leaving all item parameters fixed to the established values or use the established item parameters as priors. Using an IRT model effectively corresponds to a weighing of the observations by each item's sensitivity. Provided the ICFs fit, this weighing is optimal and will lead to an increase in statistical power.[35,40,41] Interestingly, this approach is entirely prespecifiable.[38]

These different advantages of IRT-based analyses can be applied at various points of the drug development process (e.g., we could use IRT to create a most sensitive assessment for a particular population in proof-of-concept trials, benefit from the increased sensitivity of IRT when analyzing the trial results, and finally predict the outcome of a phase III trial with a wider patient population and a regulatory accepted end point).

The increased complexity of an IRT-based analysis, mentioned at the beginning of this section, is maybe its most considerable downside. Models consist of many more components that need to be checked, temporal evolution happens on a latent scale that needs to be understood, and the highly nonlinear nature of the model needs to be handled by the parameter estimation algorithm. The IRT models also face the same bias-variance tradeoff as all other statistical models and one can argue that the flexibility of the model together with a large number of parameters also lead to a higher risk of overfitting. This overfitting risk is especially true for small datasets, in which the ICFs could be driven by outlying subjects, leading to biased results and wrong conclusions.

We can mitigate against these risks by being cautious in regard to model complexity, and we can use some of the techniques presented in this tutorial to help during model building. Those that are willing to spend this additional effort will find an IRT an extremely valuable addition to the pharmacometrics toolbox when it comes to the challenging task of quantifying a disease.

**ACKNOWLEDGMENTS.** The author thanks Elodie Plan and Mats Karlsson for many useful inputs and valuable comments to this tutorial.

**CONFLICT OF INTEREST.** The author declared no conflict of interest.

1. Bech, P., Rasmussen, N.A., Olsen, L.R., Noerholm, V. & Abildgaard, W. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *J. Affect. 6Disord.* **66**,159–164 (2001).
2. Cummings, J.L., Mega, M., Gray, K., Rosenberg-Thompson, S., Carusi, D.A. & Gornbein, J. The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. *Neurology* **44**, 2308–2314 (1994).
3. Kurtzke, J.F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* **33**, 1444–1452 (1983).
4. Bock, R.D. A brief history of item theory response. *J. Educ. Meas.* **16**, 21–33 (1997).
5. Lazarsfeld, P. The logical and mathematical foundation of latent structure analysis. 362–412 In: *Measurement and* prediction. Vol. 4 of Studies in Social Psychology in World War II. (Princeton University Press, Princeton, NJ, 1950).
6. Lord, F. A Theory of Test Scores (Psychometric Monograph No. 7). Psychometric Corporation. <http://www.psychometrika.org/journal/online/MN07.pdf> (1952).
7. Birnbaum, A. On the estimation of mental ability. (Randolph Air Force Base, TX, USAF School of Aviation Medicine, 1958).
8. Rudner, L.M. Implementing the Graduate Management Admission Test (R) Computerized Adaptive Test. (ed., Weiss, D.J.). Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing; 2007.
9. Mills, C.N. The GRE Computer Adaptive Test: Operational Issues (eds. van der Linden, W.J. & Glas, C.A.W.) 75–99 *Computerized Adaptive Testing: Theory and Practice* (Dordrecht, The Netherlands, Kluwer Academic Publishers, 2000).
10. Beal, S.L., Sheiner, L.B., Boeckman, A. & Bauer, R.J. NONMEM User's Guides (1989–2013).
11. Hu, C., Szapary, P.O., Mendelsohn, A.M. & Zhou, H. Latent variable indirect response joint modeling of a continuous and a categorical clinical endpoint. *J. Pharmacokinet. Pharmacodyn.* **41**, 335–349 (2014).
12. Reckase, M.D. Unidimensional Item Response Theory Models. 11–55 In: *Multidimensional Item Response Theory Statistics for Social and Behavioral Sciences* (New York, NY, Springer New York, 2009).
13. Sheiner, L.B. A new approach to the analysis of analgesic drug trials, illustrated with bromfenac data. *Clin. Pharmacol. Ther.* **56**, 309–322 (1994).
14. Chalmers, R.P. MIRT: a multidimensional item response theory package for the R environment. *J. Stat. Softw.* **48**, 1–29 (2012).
15. Stan: a C library for probability and sampling, version 2.8.0; 2015. http://mc-stan.org/.
16. *Stan Modeling Language Users Guide and Reference Manual*, version 2.8.0. <http://mc-stan.org/> (2015).
17. DeMars, C. *Item Response Theory* (Oxford University Press, 2010).
18. Balsis, S., Unger, A.A., Benge, J.F., Geraci, L. & Doody, R.S. Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimers Dement.* **8**, 288–294 (2012).
19. Dayneka, N.L., Garg, V. & Jusko, W.J. Comparison of four basic models of indirect pharmacodynamic responses. *J. Pharmacokinet. Biopharm.* **21**, 457–478 (1993).
20. Hastie, T. GAM: Generalized Additive Models; 2015. R package version 1.12. <https://CRAN.R-project.org/package=gam> (2015).
21. Orlando, M. & Thissen, D. Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* **24**, 50–64 (2000).
22. Orlando, M.M. & Thissen, D. Further investigation of the performance of S - X2: an item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* **27**, 289–298 (2003).
23. Gottipati, G., Karlsson, M.O. & Plan, E.L. Modeling a composite score in Parkinson's disease using item response theory. *AAPS J.* **19**, 837–845 (2017).
24. Maydeu-Olivares, A. Goodness-of-fit assessment of item response theory models. *Measurement (Mahwah N J)* **11**, 71–101 (2013).
25. Maydeu-Olivares, A. & Joe, H. Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika* **71**, 713–732 (2006).
26. Browne, M.W. & Cudeck, R. Alternative ways of assessing model fit. *Sociol. Methods Res.* **21**, 230–258 (1992).
27. Krekels, E.H., Novakovic, A.M., Vermeulen, A.M., Friberg, L.E. & Karlsson, M.O. Item response theory to quantify longitudinal placebo and paliperidone effects on PANSS scores in schizophrenia. *CPT Pharmacometrics Syst. Pharmacol.* **6**, 543–551 (2017).
28. Ueckert, S., Lockwood, P., Schwartz, P. & Riley, S. Modeling the Neuropsychiatric Inventory (NPI) - Strengths and Weaknesses of a Multidimensional Item Response Theory Approach. Abstracts for American Conference on Pharmacometrics (ACoP6). <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A915468&dswid=article> (2015).
29. Reckase, M.D. *Multidimensional Item Response Theory.* 2009 ed. (Dordrecht, The Netherlands, New York Springer, 2009).
30. Wainer, H., Bradlow, E.T. & Wang, X. *Testlet Response Theory and Its Applications.* 1st ed. (New York, NY, Cambridge University Press, 2007).
31. Germovsek, E., Ambery, C., Yang, S., Beerahee, M., Karlsson, M.O. & Plan, E.L. Handling frequent observations of composite scores: application to PROs in COPD. In: Abstracts of the Annual Meeting of the Population Approach Group in Europe. Hersonissos, Crete, Greece; 2017.
32. Sijtsma, K. & Molenaar, I.W. Introduction to Nonparametric Item Response Theory. (Thousand Oaks, CA, SAGE Publications, 2002).
33. Ueckert, S., Plan, E.L., Ito, K., Karlsson, M.O., Corrigan, B. & Hooker, A.C. Predicting baseline ADAS-cog scores from screening information using item response theory and full random effect covariate modeling. In: Abstracts for American Conference on Pharmacometrics (ACoP) 2013. Fort Lauderdale, FL, 2013.
34. Ueckert, S., Hooker, A.C., Karlsson, M.O. & Plan, E.L. Item response theory model as support for decision-making: simulation example for inclusion criteria in Alzheimer's trial. In: Abstracts of the Annual Meeting of the Population Approach Group in Europe, 2014.
35. Ueckert, S. *et al.* Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. *Pharm. Res.* **31**, 2152–2165 (2014).
36. Kalezic, A., Savic, R., Munafo, A., Plan, E.L. & Karlsson, M.O. Sample size calculations in multiple sclerosis using pharmacometrics methodology: comparison of a composite score continuous modeling and Item Response Theory approach. In: Abstracts of the Annual Meeting of the Population Approach Group in Europe, 2014.
37. Vandemeulebroecke, M., Bornkamp, B., Krahnke, T., Mielke, J., Monsch, A. & Quarg, P. A longitudinal item response theory model to characterize cognition over time in elderly subjects. *CPT Pharmacometrics Syst. Pharmacol.* **6**, 635–641 (2017).
38. Buatois, S., Retout, S., Frey, N. & Ueckert, S. Item response theory as an efficient tool to describe a heterogeneous clinical rating scale in de novo idiopathic Parkinson's disease patients. *Pharm. Res.* **34**, 2109–2118 (2017).
39. Schindler, E. *et al.* Analyzing patient-reported outcomes in breast cancer through item-response theory pharmacometric modeling. In: Abstracts for American Conference on Pharmacometrics (ACoP6) (2015).
40. Schindler, E., Friberg, L.E. & Karlsson, M.O. Comparison of item response theory and classical test theory for power/sample size for questionnaire data with various degrees of variability in items' discrimination parameters. In: Abstracts of the Annual Meeting of the Population Approach Group in Europe (2015).
41. Vlitalo, P.A.J., Rottschfer, V., van Dijk, M., Knibbe, C.A.J. & Krekels, E.H.J. A mathematical method assessing the informativeness of total test score on the basis of an item response theory model: a case study on pain in children. In: Abstracts of the Annual Meeting of the Population Approach Group in Europe (2016).

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (http://psp-journal.com)