Supervised Multidimensional Item Response Theory Modeling of Pediatric Iatrogenic Withdrawal Symptoms

Goulooze, S.C.; Ista, E.; Dijk, M. van; Hankemeier, T.; Tibboel, D.; Knibbe, C.A.J.; Krekels, E.H.J.

Citation: CPT Pharmacometrics Syst. Pharmacol. (2019) 8, 904–912; doi:10.1002/psp4.12469

ARTICLE

Supervised Multidimensional Item Response Theory Modeling of Pediatric Iatrogenic Withdrawal Symptoms

Sebastiaan C. Goulooze[1], Erwin Ista[2], Monique van Dijk[2], Thomas Hankemeier[1], Dick Tibboel[2], Catherijne A.J. Knibbe[1,3] and Elke H.J. Krekels[1,*]

Item-level data from composite scales can be analyzed with pharmacometric item response theory (IRT) models to improve the quantification of disease severity compared with the use of total composite scores. However, regular IRT models assume unidimensionality, which is violated in the scale measuring iatrogenic withdrawal in children because some items are also affected by pain, undersedation, or delirium. Here, we compare regular IRT modelling of pediatric iatrogenic withdrawal symptom data with two new analysis approaches in which the latent variable is guided towards the condition of interest using numerical withdrawal severity scored by nurses as a "supervising variable:" supervised IRT (sIRT) and supervised multi-dimensional (smIRT) modelling. In this example, in which the items scores are affected by multiple conditions, regular IRT modeling is worse to quantify disease severity than the total composite score, whereas improved performance compared with the composite score is observed for the sIRT and smIRT models.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?
☑ The item-level data from composite scales can be analyzed with pharmacometric item response theory (IRT) models, which improves the quantification of the disease severity compared with the total composite score. However, regular IRT models assume unidimensionality, which is not the case for clinical scales of iatrogenic withdrawal because its items can also be affected by pain, undersedation, or delirium.

WHAT QUESTION DID THIS STUDY ADDRESS?
☑ The study analyzes data from a composite scale for monitoring pediatric iatrogenic withdrawal and compares the performance of regular IRT modeling and the following two novel methods of IRT modeling: supervised IRT (sIRT) and supervised multidimensional IRT (smIRT).

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
☑ Regular IRT modeling was worse than the total composite score in quantifying iatrogenic withdrawal. Both sIRT and smIRT modeling provided superior quantification of withdrawal compared with the total composite score.

HOW MIGHT THIS CHANGE DRUG DISCOVERY, DEVELOPMENT, AND/OR THERAPEUTICS?
☑ The supervised IRT methods introduced here can improve the analysis of data from composite clinical scales that are not unidimensional.

In the pediatric intensive care unit (ICU), critically ill children often receive opioids and sedatives for prolonged periods of time, which may contribute to drug dependence and iatrogenic withdrawal syndrome (IWS).[1,2] The importance of careful weaning from these drugs, instead of abrupt discontinuation, is well established. However, even in studies with standardized weaning protocols, IWS is reported in 5–87% of children.[1,3] Lacking strategies for predicting individualized weaning strategies, weaning is guided by the frequent monitoring for IWS.[4] However, the lack of specific withdrawal symptoms makes monitoring for IWS difficult because a particular symptomatic profile (e.g., a crying child with tense body parts, anxiety, and hyperalertness) might be caused by IWS but also by pain, undersedation, or delirium as illustrated in **Figure 1**.[5–7] In practice, ICU nurses use contextual information to decide on the most likely explanation of the symptoms.[5,8] For example, IWS might be expected during weaning from opioids.

The Sophia Observational withdrawal Scale ($SOS_{withdrawal}$) is a validated scale that scores the presence of 15 symptoms related to IWS.[9,10] The number of symptoms present equals the composite score of this scale, which is used to identify IWS. As such, all symptoms are weighed equally and treated as equally informative of IWS. However, if some symptoms are more informative of IWS than other symptoms, a superior quantification of IWS will be obtained by

[1]Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands; [2]Intensive Care and Pediatric Surgery, Erasmus Medical Center-Sophia Children's Hospital, Rotterdam, The Netherlands; [3]Department of Clinical Pharmacy, St. Antonius Hospital, Nieuwegein, The Netherlands. *Correspondence: Elke H.J. Krekels (e.krekels@lacdr.leidenuniv.nl)
Received: May 2, 2019; accepted: August 14, 2019. doi:10.1002/psp4.12469

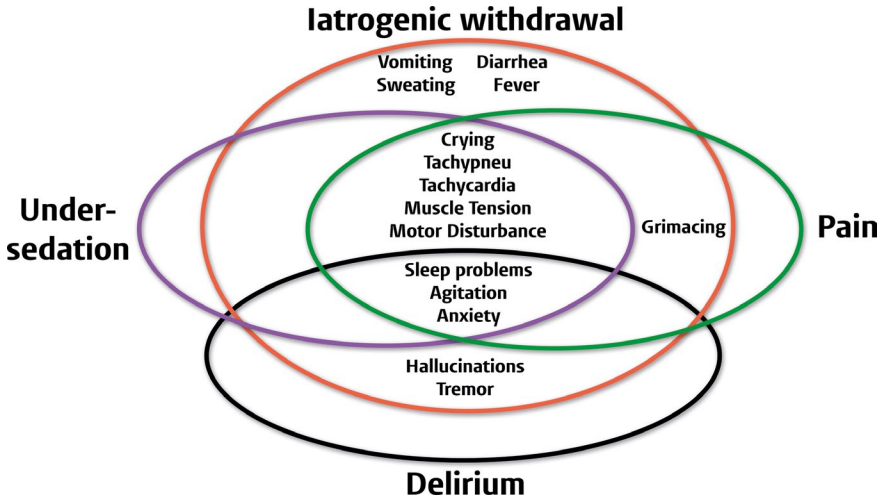

Figure 1 Overview of iatrogenic withdrawal syndrome–associated symptoms included in Sophia Observational withdrawal Scale assessment and their suggested overlap with pain, undersedation, and delirium in the 2016 position statement by the European Society of Paedetric and Neonatal Intensive Care (ESPNIC).[5,9]

weighing symptoms according to their informativeness. This can be done using item response theory (IRT) models, which use the observed item scores to estimate the unobserved or hidden latent variable, which is generally regarded in pharmacometrics as an approximation of the disease severity.[11–13] Such IRT-based analyses have been reported to have improved statistical power and precision over analyses that regard the total scores of composite scales.[13]

However, IRT models rely on unidimensionality, the assumption that a single shared factor influences the probability of each of the withdrawal symptoms.[11] This assumption is violated in the $SOS_{withdrawal}$ because IWS, pain, undersedation, and delirium can all cause certain symptoms on the scale, and children in the ICU can suffer from multiples of these conditions simultaneously. When using an unsupervised technique such as IRT modeling, essentially a single latent variable is identified that captures as much of the variability and correlations within the item-level score data as possible.[14–16] However, considering that pain, undersedation, and delirium are ubiquitous in the pediatric ICU, there is no guarantee that the latent variable in a regular IRT analysis of the $SOS_{withdrawal}$ is a good approximation of IWS.

Recently, an extension of the regular IRT, which takes this technique from an unsupervised to a supervised implementation, was proposed by Idé and Dhurandhar[14] in a machine-learning context. By using an external outcome or supervising variable, a supervised IRT (sIRT) model is directed toward a latent variable that represents the condition of interest rather than the latent variable that captures most of the variability in the item-level score data. However, by guiding the latent variable toward IWS, we might fail to account for some correlations in the withdrawal symptom data. This could result in a violation of the conditional independence assumption when, for example, symptoms associated with pain are correlated independent of IWS. If this does indeed result in misspecification, the sIRT model may be extended toward a multidimensional setting by including one or more additional latent variables in the model to capture remaining correlation patterns.[11,17]

In this study, we reanalyzed data from the clinical study that validated the $SOS_{withdrawal}$ for IWS assessment in critically ill children.[9,10] The first aim was to investigate the ability of regular IRT modeling quantify withdrawal severity in a composite scale in which the unidimensionality assumption may be violated. The second aim was to investigate if the withdrawal severity could be improved by extending the regular IRT model to a sIRT model or a supervised multidimensional IRT model (smIRT).

METHODS

Clinical study

Data were collected during an observational study by Ista *et al.*,[10] in which the $SOS_{withdrawal}$ was validated for IWS assessment in children. The study was approved by the local institutional review board, which waived the need for parental informed consent. The study included 154 children who received continuous infusions of midazolam, morphine, or fentanyl for 5 or more days. To obtain a more homogenous population, we excluded data from neonates and children who received extracorporeal membrane oxygenation therapy. In the 81 children analyzed in this work, a total of 1,785 $SOS_{withdrawal}$ assessments were collected.

The $SOS_{withdrawal}$ assessment scores the occurrence of the 15 withdrawal-associated symptoms (**Figure 1**) with 13 binary items and two items with more than two possible outcomes: tremors and motor disturbance. Because of the low incidence of tremors in the data set, this data item was also treated as a binary item to increase model stability. For the motor disturbance item, scores of 0, 1, and 2 are possible. The same nurse who performed the $SOS_{withdrawal}$ assessment also scored IWS severity on a 0–10 numerical rating scale ($NRS_{withdrawal}$). When scoring the $NRS_{withdrawal}$, contextual factors are taken into account to better distinguish IWS from conditions with similar symptomatic profiles. The observed item-pair correlations in the $SOS_{withdrawal}$ data are shown in **Figure S1**.

Model development

We developed the following three different models: a regular IRT model, a sIRT model, and a smIRT. Model parameters were estimated using the stochastic approximation expectation maximization estimation method in NONMEM 7.3 (ICON, Dublin, Ireland), a general-purpose software package for maximum likelihood parameter estimation of nonlinear mixed-effects models.[18] We obtained the objective function value (OFV) and covariance matrix by performing the expectation step from the importance sampling method with the final parameter estimates.

For all three IRT models, parametric item characteristic curves (ICC) were used to describe the relationship between the latent variable and the probability of observing a particular score on a specific item. For the binary items, the following ICC was used:

$$P(Y_{ij}=1)=c_j\cdot\frac{e^{a_j\cdot(LV-b_j)}}{1+e^{a_j\cdot(LV-b_j)}} \quad (1)$$

where c_j represents the maximum probability of observing a particular symptom, a_j is the item-specific discrimination parameter, b_j represents the item-specific difficulty parameter, and LV is the latent variable for iatrogenic withdrawal from either the regular IRT model or the two sIRT models. For the item motor disturbance, a two-parameter logit model was used, as described by Ueckert.[11]

The development of both the regular IRT model and the sIRT model started with a base model consisting of two-parameter logit ICCs for all items, which is obtained when c_j in Eq. 1 is fixed to 1. ICCs with an estimated value for c_j were considered a significant improvement ($P < 0.001$) over the base ICC if this resulted in a drop in OFV of more than 10.8 points.

Regular IRT modeling

In regular IRT modeling, the latent variable consists of a random effects term on an arbitrary scale between $-\infty$ and $+\infty$:

$$LV_{IRT,i}=\eta_i \quad (2)$$

where η represents the random effects parameter that is normally distributed with an estimated mean and an estimated variance of ω^2. To maintain identifiability of the model parameters, we fixed the ICC of the tachycardia item to equal that of the sIRT model described later. This was done so that the ICCs of the regular IRT model could be visually compared with that of the sIRT model.

sIRT modeling

In the sIRT model, the single latent variable is conditioned on an external outcome or "supervising" variable. In this study, the supervising variable was the $NRS_{withdrawal}$ score described previously.

To maintain the bounded nature of the NRS score, random effects of the latent variable were added on a logit scale before transforming back to a 0–10 scale, as shown in Eqs. 3–5.

$$NRS_{transform,i}=\frac{NRS_{withdrawal,i}}{10}\cdot(1-\delta)+\frac{\delta}{2} \quad (3)$$

$$LV_{logit,i}=\log\left(\frac{NRS_{transform,i}}{1-NRS_{transform,i}}\right)+\eta_i \quad (4)$$

$$LV_{sIRT,i}=10\cdot\frac{e^{LVlogit,i}}{1+e^{LVlogit,i}} \quad (5)$$

where δ in Eq. 3 represents a small constant used to allow the logit transformation of $NRS_{withdrawal}$ in the presence of data on the boundaries of the NRS scale. The δ was set arbitrarily to 0.02 after a sensitivity analysis revealed that there was a negligible impact (<10%) on all the estimated parameters of the model when varying this value between 0.05 and 0.0001 (data not shown). η represents the random effects parameter that is normally distributed with a mean of zero and an estimated variance of ω^2.

smIRT modeling

The sIRT model with a single latent variable was extended to a multidimensional setting by adding one or two additional latent variables to the model. This was done to reduce the violation of the local independence assumption by accounting for factors other than IWS that can affect the items of the $SOS_{withdrawal}$ scale. Here, we used a compensatory multidimensional IRT model extension of Eq. 5:

$$P(Y_{ij}=1)=c_j\cdot\frac{e^{a1_j\cdot LV_1+a2_j\cdot LV_2+a3_j\cdot LV_3-b_j}}{1+e^{a1_j\cdot LV_1+a2_j\cdot LV_2+a3_j\cdot LV_3-b_j}} \quad (6)$$

where LV_1 represents the first, supervised latent variable, and LV_2 and LV_3 represent the additional second and third latent variables. $a1_j$, $a2_j$, and $a3_j$ represent the item-specific discrimination parameters for the first, second, and third latent variables, respectively; b_j represents the item-specific difficulty parameter.[19] A similar adjustment was made to the two-parameter logit model that was used for the motor disturbance item as described previously. The distributions of additional latent variables in the current data set were set to a standard normal distribution (see Eq. 2) to allow identifiability of the model. The first latent variable was defined as the sIRT model with estimated random effects parameter around the $NRS_{withdrawal}$ score (Eq. 3).

Initially, we associated additional latent variables with all 15 items, estimating a nonzero a_j parameter for each item. We also explored an alternative strategy, where the impact of additional latent variables on the item probabilities was only estimated for a subset of the items by fixing $a2_j$ or $a3_j$ to zero for the other items. The subset of items affected by the additional latent variables was selected based on prior clinical knowledge about which items are affected by conditions other than IWS (**Figure 1**), the correlation matrix of the standardized residuals of the sIRT model (**Figure 2**), and the standard error of the estimates of $a2_j$ and $a3_j$ (**Table S1**).

Model evaluation and comparison between models

In models with a single latent variable, the item-specific fit of the ICC of each item were evaluated by visual

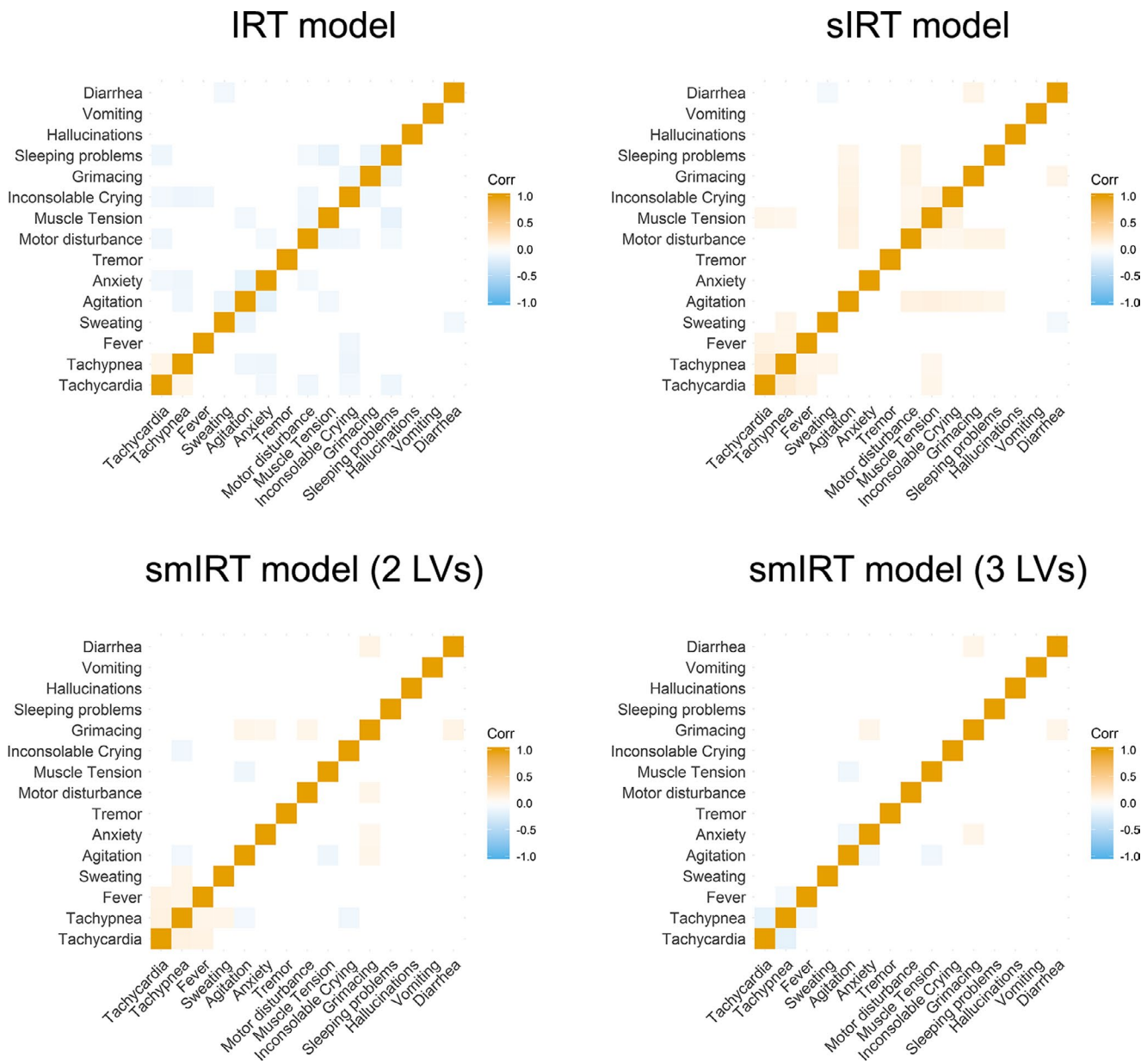

Figure 2 Heat map of the statistically significant ($P < 0.001$) correlations of standardized residuals between different items of the following four different model types: regular IRT, sIRT, smIRT with two or three LVs. Significant correlations between item pairs indicate that these items are not independent when conditioned on the model. IRT, item response theory; LV, latent variable; sIRT, supervised IRT model; smIRT, supervised multidimensional model.

comparison of the estimated ICCs from the model with a generalized additive model–based nonparametric smoother of the ICC.[11,13] The estimated ICCs of the sIRT model and the regular IRT model were also visually compared with each other.

NONMEM's covariance step was used to determine the relative standard error of the parameter estimates. To assess collinearity in the model parameters, we calculated the condition number as the square root of the ratio between the largest and smallest eigenvalue of the correlation matrix. Condition numbers above 20 were considered to suggest overparameterization.[20] The conditional independence was evaluated by examining a heat map of the correlation matrix of the item-specific standardized residuals.[11]

An eightfold cross-validation procedure was performed to evaluate the linear association between the $NRS_{withdrawal}$ score and the *post hoc* estimate of the latent variable of the regular IRT, sIRT, and smIRT models. To allow a direct comparison of the regular and sIRT models, the *post hoc* estimation step for all models was done in the absence of the $NRS_{withdrawal}$ score by replacing the $NRS_{withdrawal}$ score in Eq. 3 with an estimated median latent variable of the population while fixing the ICC curves estimated in the presence of the $NRS_{withdrawal}$. The linear association was determined by calculating the Akaike information criterion (AIC) of linear

models in which the latent variable of a particular IRT model was the predictor and the $NRS_{withdrawal}$ score was the dependent variable. As a benchmark, we also examined the AIC of a linear model in which the total SOS score was used a predictor for the $NRS_{withdrawal}$ score. A lower AIC indicates a stronger linear association with the $NRS_{withdrawal}$ score. Details on the cross-validation procedure can be found in the **Supplemental Information**.

RESULTS

Model development

Regular IRT modeling. Estimating a maximum probability parameter (c_j in Eq. 1) different from 1, did not significantly improve the regular IRT model ($P > 0.001$). Visual comparison with the nonparametric ICC showed model misspecification in most items (**Figure S2**). The condition number of 48.7 exceeded the threshold value of 20. A heat map of the correlation matrix of the standardized residuals is shown in **Figure 2**. Many items have (mostly negatively) correlated residuals ($P < 0.001$), but these items do not appear to form clusters of correlated items.

sIRT model. In the final sIRT model with the $NRS_{withdrawal}$ as a supervising variable, a maximum probability of observing a symptom was estimated on five items, i.e., agitation, inconsolable crying, grimacing, sleeping problems, diarrhea. When compared with the regular IRT model, the sIRT model had a good item-specific fit of the data in most items (**Figure S3**). The condition number of 14 indicated a limited degree of collinearity (**Table 1**). A heat map of the correlation matrix of the standardized residuals is shown in **Figure 2**. In contrast with the regular IRT model, the sIRT model does not have a large number of item pairs with significantly negative correlated residuals. However, there seems to be a cluster of positively correlated residuals with items agitation, motor disturbance, muscle tension, inconsolable crying, grimacing, and sleeping problems, and another cluster with the items tachycardia, tachypnea, fever, and sweating.

smIRT model. To account for the residual positive correlations in the sIRT model with one latent variable, the model was extended to a multidimensional setting by adding a second latent variable for all 15 items, which adds an additional 15 estimated parameters to the model. Because this model did not converge successfully, in an adapted approach the second latent variable was added to only those $SOS_{withdrawal}$ items that are suggested to be associated with pain or undersedation.[5] This decision was also supported by **Figure 2**, as this figure showed that these items in particular seemed to violate the local independence assumption in the sIRT model. These items are the following: agitation, motor disturbance, muscle tension, inconsolable crying, grimacing, sleeping problems, anxiety, tachycardia, and tachypnea.[5] This smIRT model minimized successfully and results in a drop in OFV of 130.1 points with nine additional estimated parameters ($P < 0.001$). Because the relative standard error (RSE) of the estimate of the $a2_j$ parameters for the anxiety and grimacing items were high (>50%), we explored fixing $a2_j$ to zero for both items, which increased the OFV by only 7.9 points ($P > 0.001$). This was considered the best smIRT model with two latent variables. The correlation matrix of the standardized residuals depicted in **Figure 2** does not show a cluster of correlated residuals with items agitation, motor disturbance, muscle tension, inconsolable crying, grimacing, and sleeping problems, although slight correlations between the tachycardia, tachypnea, fever, and sweating items are still present for the smIRT model with two latent variables.

A smIRT model with three latent variables was also developed in which the first latent variable was a supervised latent variable to characterize IWS severity, the second

Table 1 Numerical overview of final model fits

	Regular IRT	sIRT	smIRT (two latent variables)	smIRT (three latent variables)
# estimated parameters	30	36	43	46
Condition number	48.7	14.0	16.2	24.3
Items affected by LV1	All 15 $SOS_{withdrawal}$ items	All 15 $SOS_{withdrawal}$ items	All 15 $SOS_{withdrawal}$ items	All 15 $SOS_{withdrawal}$ items
Items affected by LV2	–	–	Agitation Motor Disturbance Muscle Tension Inconsolable Crying Sleeping Problems Tachycardia Tachypnea	Agitation Motor Disturbance Muscle Tension Inconsolable Crying Sleeping Problems Grimacing
Items affected by LV3	–	–	–	Tachycardia Tachypnea Fever Sweating
OFV (with $NRS_{withdrawal}$ score)	–	17474.06	17351.82	17259.17
OFV (without $NRS_{withdrawal}$ score)[a]	18569.09	18701.1	18635.28	18590.70

IRT, item response theory; LV_1, the first, supervised latent variable; LV_2, second latent variable; LV_3, third latent variable; $NRS_{withdrawal}$, numerical rating scale score of withdrawal severity; OFV, objective function value; sIRT, supervised IRT model; smIRT, supervised multidimensional model; $SOS_{withdrawal}$, Sophia Observational withdrawal Scale.

[a]When estimating the distribution of the latent variable of the sIRT and smIRT models in the absence of the $NRS_{withdrawal}$ score, all item characteristic curves are fixed to the parameter estimates obtained during the model fit with the $NRS_{withdrawal}$ score.

latent variable was to characterize the remaining correlations on the behavioral items associated with pain or undersedation—agitation, motor disturbance, muscle tension, inconsolable crying, grimacing, sleeping problems, and anxiety—and the third latent variable to characterize the remaining correlations in the sIRT model in the items associated with autonomic dysfunction—tachycardia, tachypnea, fever, and sweating. Because the RSE of the $a2_j$ parameters for the anxiety and grimacing items were high (>50%), we explored fixing $a2_j$ to zero for both items. This increased the OFV by 6.0 points ($P > 0.001$) for anxiety and by 11.9 points ($P < 0.001$) for grimacing. Therefore, in the final smIRT model with three latent variables, $a2_j$ was estimated for grimacing, but not for anxiety. For this model, no clustered correlations among the standardized residuals of the items were observed (**Figure 2**). With a condition number of 24.3, this model had a degree of parameter collinearity slightly above the predefined threshold of 20 (**Table 1**).

Comparison of three IRT model types

The parameter estimates of the three IRT models are shown in **Table S1**. The ICCs of the regular IRT and sIRT models are shown in **Figure 3**. In addition to the item tachycardia, for which the ICC of the regular IRT model was fixed to that of the sIRT model, several other items also show similar ICCs. The items where the ICCs of the two models diverge include the item muscle tension as well as the five items for which the sIRT allowed for the estimation of a maximum probability for observing items, i.e., agitation, inconsolable crying, grimacing, sleeping problems, and diarrhea. In the absence of the NRS scores, despite having the lowest number of estimated parameters, the regular IRT model

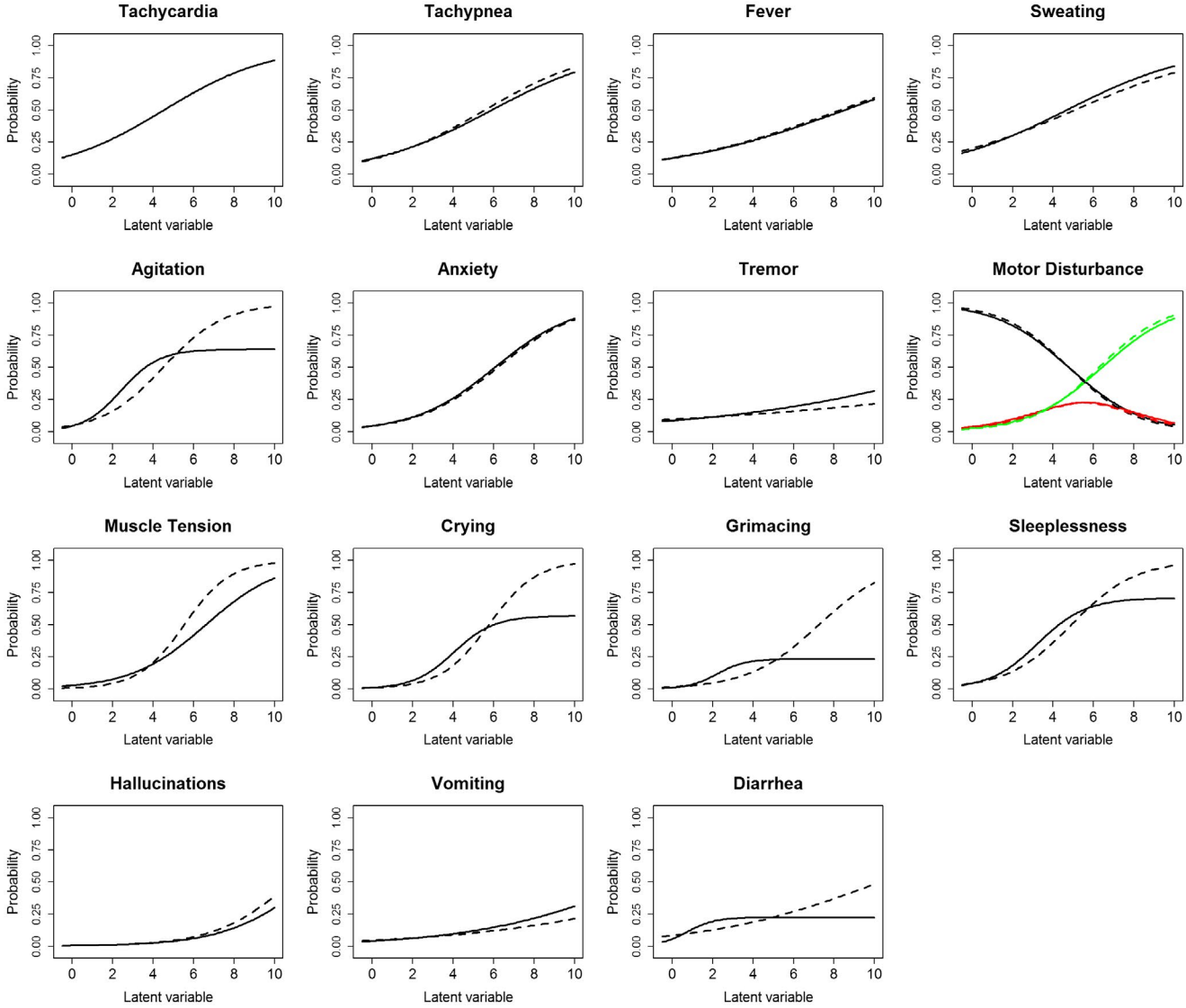

Figure 3 Comparison on the item characteristic curves of the final supervised IRT model (sIRT) model (solid lines) and the final regular item response theory (IRT) model (dashed lines). For the motor disturbance item, the probability of a score of 0 (black), 1 (red), or 2 (green) is depicted. For all other items, the probability of a score of 1 is shown. To allow visual comparison, the item characteristic curve of the tachycardia item in the regular IRT model was fixed to the final estimate of the sIRT model.

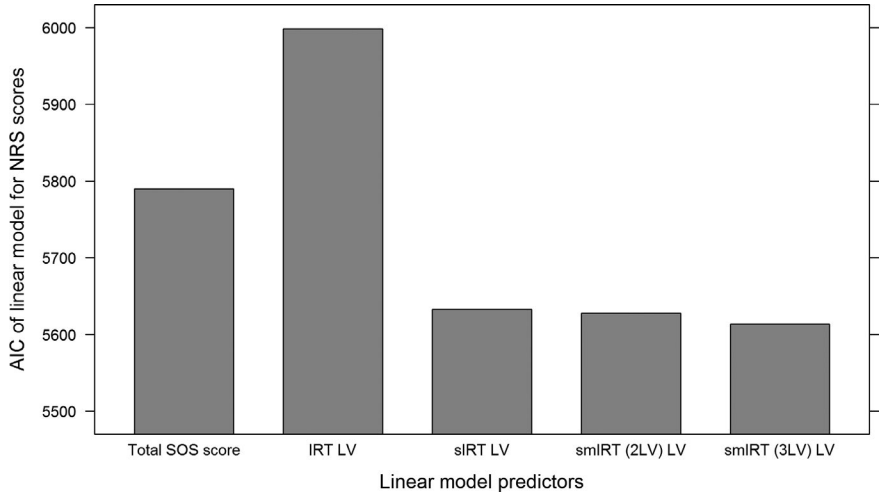

Figure 4 Comparison of association of LV of the four IRT model types and the total score of the $SOS_{withdrawal}$ scale with the $NRS_{withdrawal}$ scores. In all cases, the LVs were estimated in the absence of the $NRS_{withdrawal}$ score. AIC, Akaike information criterion; IRT, item response theory; LV, latent variable; $NRS_{withdrawal}$, numerical rating scale score of withdrawal severity by nurse; sIRT, supervised IRT model; smIRT, supervised multidimensional model; SOS, Sophia Observational withdrawal Scale.

was observed to best fit (i.e., lowest OFV) the item-level $SOS_{withdrawal}$ data (**Table 1**), meaning that this best captures the variability and correlations within the item-level score.

Figure 4 shows the results of the cross-validation procedure to assess the linear association between the $NRS_{withdrawal}$ score and the latent variables obtained with the three IRT model types in the absence of the $NRS_{withdrawal}$ score. The figure shows that the latent variable of the regular IRT showed the weakest association with the $NRS_{withdrawal}$ score (AIC = 5998.4), with an even higher AIC than the total $SOS_{withdrawal}$ score (AIC = 5789.8). This means that although most of the variability and correlations are captured by the latent variable of this model, as indicated by the lowest OFV, the latent variable does not reflect withdrawal specifically. Even after removing the $NRS_{withdrawal}$, all sIRT models had a stronger association than the total $SOS_{withdrawal}$ score, with the strongest association being observed for the first latent variable of the smIRT with three latent variables (AIC = 5613.3).

DISCUSSION

We used three different IRT modeling methods to analyze data from a composite scale validated to monitor IWS in children, i.e., $SOS_{withdrawal}$. A regular IRT model was developed to assess its performance in a practical situation where the unidimensionality assumption is known to be violated. We used the sIRT modeling approach suggested by Idé and Dhurandhar[14] to guide the latent variable toward IWS using the nurse's expert opinion that also takes contextual factors into account (i.e., $NRS_{withdrawal}$). To diminish the violation of the conditional independence assumption and better capture the correlations between the items, we extended the sIRT methodology to a multidimensional setting using both data-driven arguments and clinical knowledge during model development.

Our results demonstrate that regular IRT modeling of withdrawal symptoms might provide the lowest OFV when compared with sIRT models in the absence of the NRS score (**Table 1**), but that considerable misspecification was observed in the item-specific fit (**Figure S2**). This misspecification might explain the large number of negatively correlated residuals between item pairs (**Figure 2**). More important, the latent variable in the regular IRT model does not provide a good approximation of IWS severity, as it has a weaker association to the $NRS_{withdrawal}$ than the total score of the $SOS_{withdrawal\ scale}$ (**Figure 4**). This suggests that the latent variable of the regular IRT model does not selectively quantify IWS, but likely a mixture of withdrawal and related conditions such as pain, undersedation, and delirium. In such situations, it is unlikely that regular IRT modeling of withdrawal symptoms will have improved statistical power compared with approaches that model total scores, which is an important argument to use IRT modeling in pharmacometrics.[11,13] It might even lead to erroneous conclusions when using this regular IRT model as a basis for developing a longitudinal pharmacometric model. Finally, redesigning the clinical scale by removing items that are not informative in the regular IRT model might also be counterproductive when the model does not selectively model the condition of interest.

With the sIRT models, additional information was used to guide the latent variable toward the condition of interest. In this study, the $NRS_{withdrawal}$ score given by trained nurses was used for this purpose during the estimation of the ICCs. When the $NRS_{withdrawal}$ scores were removed from the data set and the latent variable reestimated, we found a stronger association of the latent variable with the $NRS_{withdrawal}$ score than the total score in all sIRT models (**Figure 4**). With the sIRT models, we also encountered less problems with model convergence than with the regular IRT. This might be explained by the elevated collinearity in the parameter estimates of the regular IRT model (**Table 1**). Finally, the sIRT showed improved item-specific fit of the data (**Figure S3**), which might also explain why there were less item pairs with negatively correlated residuals compared with the regular

IRT model (**Figure 2**). However, the sIRT model had clusters of items with positively correlated residuals, which were not observed in the regular IRT model.

We extended the sIRT model to a multidimensional setting to account for these residual correlations. These residual correlations indicate a violation of the conditional independence assumption and might be caused by the fact that as we are "guiding" the model to focus on IWS more selectively in the sIRT model compared with the regular IRT model so that residual correlations could emerge between items affected by other conditions such as pain, undersedation, and delirium. Informed by clinical knowledge, we included an effect of additional latent variables on groups of specific items (see **Table 1**) that are affected by similar underlying conditions (**Figure 1**).[5] Although the smIRT model with three latent variables had a stronger association with the $NRS_{withdrawal}$ score than its unidimensional counterpart, this modest difference is unlikely to indicate a meaningfully improved predictive performance (**Figure 4**).

Similar to regular IRT models, the supervised models developed here can be extended to longitudinal models so that IWS can be modeled as a function of time, drug concentration, or other predictors. For modeling $SOS_{withdrawal}$ item-level data in children for which the $NRS_{withdrawal}$ score is unavailable, it could be appealing to use the supervised models presented here to estimate the latent variable and then model the latent variable as a continuous outcome variable.[21] Considering that the latent variable of all sIRT models had a stronger association with the $NRS_{withdrawal}$ scores than the total $SOS_{withdrawal}$ score, this might improve the statistical power of such an analysis.

The interpretation of clinical composite scale data is difficult in situations where a number of separate items show a lack of specificity for the condition of interest, for example, most items of the $SOS_{withdrawal}$ scale are not specific to IWS. The sIRT models presented here can improve the statistical power and the interpretability when compared with regular IRT models of such data. An important part of the sIRT model development is the choice for "supervising variable." Depending on the goal of the analysis and availability, different types of data might be considered as "supervising variables" to guide the latent variable toward the condition of interest, such as overall quality-of-life scores, severity scores by clinical experts, or clinical end points such as survival.

The $NRS_{withdrawal}$ score is a suitable supervising variable that provided additional information on the context of the observations. However, the $NRS_{withdrawal}$ is more subjective than $SOS_{withdrawal}$ and depends more strongly on the experience of the nurse, which complicates its implementation in standardized treatment protocols. In practice it is beneficial to combine the $SOS_{withdrawal}$ with expert opinion (i.e., $NRS_{withdrawal}$). This approach combines objective symptomatic observations with the nurse's knowledge of contextual information.[5] With the sIRT and smIRT models, we improve the information obtained from the symptom data beyond the total $SOS_{withdrawal}$ score, even in the absence of the $NRS_{withdrawal}$ (**Figure 4**). Using the models developed here to estimate the IWS severity from symptom data alone can therefore be useful when $NRS_{withdrawal}$ scores are lacking or as an objective supplement to the subjective $NRS_{withdrawal}$ score.

In summary, for clinical composite scales such as the $SOS_{withdrawal}$ in which individual items may be affected by conditions other than the condition of interest, regular IRT modeling might be worse in terms of quantifying disease severity than analysis approaches based on total score. This is markedly improved when using sIRT in which the latent variable is "guided" toward the condition of interest using additional information as a supervising variable. Further improvement can be achieved by dealing with violations of the conditional independence assumption by adding a multidimensional component to the model with additional latent variables.

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

Figure S1.
Figure S2.
Figure S3.
Table S1.
Model Code sIRT. sIRT, supervised item response theory model.
Model Code smIRT. smIRT, supervised multidimensional item response theory extension model.
Supplemental Methods.

Funding. C.K. is supported by The Netherlands Organisation for Scientific Research (NWO) through a personal Vidi grant (Knibbe, 2013).

Conflict of Interest. All authors declared no competing interests for this work.

Author Contributions. S.C.G. and E.H.J.K. wrote the manuscript. E.I., M.D., and D.T. performed the research. S.C.G., E.I., M.D., C.A.J.K., E.H.J.K., and T.H. designed the research. S.C.G. analyzed the data.

1. Best, K.M., Boullata, J.I. & Curley, M.A. Risk factors associated with iatrogenic opioid and benzodiazepine withdrawal in critically ill pediatric patients: a systematic review and conceptual model. *Pediatr. Crit Care Med.* **16**, 175–183 (2015).
2. Tobias, J.D. Tolerance, withdrawal, and physical dependency after long-term sedation and analgesia of children in the pediatric intensive care unit. *Crit. Care Med.* **28**, 2122–2132 (2000).
3. Franck, L.S., Naughton, I. & Winter, I. Opioid and benzodiazepine withdrawal symptoms in paediatric intensive care patients. *Intensive Crit. Care Nurs.* **20**, 344–351 (2004).
4. Ista, E. & van Dijk, M. Knowing risk factors for iatrogenic withdrawal syndrome in children may still leave us empty-handed. *Crit. Care Med.* **45**, 141–142 (2017).
5. Harris, J. *et al.* Clinical recommendations for pain, sedation, withdrawal and delirium assessment in critically ill infants and children: an ESPNIC position statement for healthcare professionals. *Intensive Care Med.* **42**, 972–986 (2016).
6. Cramton, R.E. & Gruchala, N.E. Babies breaking bad: neonatal and iatrogenic withdrawal syndromes. *Curr. Opin. Pediatr.* **25**, 532–542 (2013).
7. Madden, K., Burns, M.M. & Tasker, R.C. Differentiating delirium from sedative/hypnotic-related iatrogenic withdrawal syndrome: lack of specificity in pediatric critical care assessment tools. *Pediatr. Crit Care Med.* **18**, 580–588 (2017).
8. Ista, E. & van Dijk, M. We cannot compartmentalize our patients! Overlapping symptoms of iatrogenic withdrawal syndrome, pediatric delirium, and anticholinergic toxidrome. *Pediatr. Crit Care Med.* **18**, 603–604 (2017).
9. Ista, E., van Dijk, M., de Hoog, M., Tibboel, D. & Duivenvoorden, H.J. Construction of the Sophia Observation withdrawal Symptoms-scale (SOS) for critically ill children. *Intensive Care Med.* **35**, 1075–1081 (2009).
10. Ista, E., de Hoog, M., Tibboel, D., Duivenvoorden, H.J. & van Dijk, M. Psychometric evaluation of the Sophia Observation withdrawal symptoms scale in critically ill children. *Pediatr. Crit Care Med.* **14**, 761–769 (2013).
11. Ueckert, S. Modeling composite assessment data using item response theory. *CPT. Pharmacometrics. Syst. Pharmacol.* **7**, 205–218 (2018).

12. Valitalo, P.A. *et al.* Pain and distress caused by endotracheal suctioning in neonates is better quantified by behavioural than physiological items: a comparison based on item response theory modelling. *Pain* **157**, 1611–1617 (2016).
13. Ueckert, S., Plan, E.L., Ito, K., Karlsson, M.O., Corrigan, B. & Hooker, A.C. Improved utilization of ADAS-cog assessment data through item response theory based pharmacometric modeling. *Pharm. Res.* **31**, 2152–2165 (2014).
14. Idé, T. & Dhurandhar, A. Supervised item response models for informative prediction. *Knowl. Inf. Syst.* **51**, 235–257 (2017).
15. Cuperlovic-Culf, M. Machine learning methods for analysis of metabolic data and metabolic pathway modeling. *Metabolites* **8**, 1–16 (2018).
16. Greene, C.S., Tan, J., Ung, M., Moore, J.H. & Cheng, C. Big data bioinformatics. *J. Cell. Physiol.* **229**, 1896–1900 (2014).
17. Ueckert, S., Lockwood, P., Schwartz, P. & Riley, S. Modeling the neuropsychiatric inventory (NPI) strengths and weaknesses of a multidimensional item response theory approach. *J. Pharmacokinet Pharmacodyn.* **42**, S92 (2015).
18. Beal, S.L. & Sheiner, L.B. NONMEM User's Guides (NONMEM Project Group, University of California, San Francisco, CA, 1992).
19. Reckase, M.D. Multidimensional Item Response Theory Models. Multidimensional Item Response Theory (Springer, New York, 2009).
20. Mould, D.R. & Upton, R.N. Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods. *CPT. Pharmacometrics. Syst. Pharmacol.* **2**, 1–14 (2013).
21. Schindler, E. *et al.* A pharmacometric analysis of patient-reported outcomes in breast cancer patients through item response theory. *Pharm. Res.* **35**, 1–14 (2018).