Science Education
Review Letters

Using polytomous IRT models to evaluate theoretical levels of understanding models and modeling in biology education

Moritz Krell
Freie Universität Berlin, Berlin, Germany

Moritz.Krell@fu-berlin.de

Received 4th June 2012, Accepted 2nd October 2012

Abstract
Forced choice-tasks have been developed to assess students' (*N* = 901) understanding of models and modeling in biology based on a theoretical structure differentiating five aspects and three levels of understanding (Upmeier zu Belzen & Krüger, 2010). The data have been analyzed by using the partial credit and the rating scale model to evaluate the assumption of an increasing degree of difficulty from level I to level III in each aspect. The findings suggest (1) that the levels of understanding are not equidistant across all aspects and (2) that the theoretically developed levels of understanding are in fact ordered by difficulty. However the latter issue needs further investigations since the present findings are not clear for all five aspects of the theoretical structure.

Keywords
polytomous IRT models, partial credit model, rating scale model, model competence, levels of understanding, biology education

Introduction

The importance of models and modeling for science education has been highly recognized (e.g. Oh & Oh, 2011; Windschitl, Thompson, & Braaten, 2008). Upmeier zu Belzen and Krüger (2010) have developed a theoretical structure of understanding models and modeling in biology education with the five aspects *nature of models, multiple models, purpose of models, testing models*, and *changing models*. Each aspect is further differentiated into three levels of understanding which reflect an increasing degree of complexity. Such a structure may be used as an analytic framework to investigate students' understanding of models and modeling and the success of teaching units in a detailed way. However the theoretical structure is still hypothetical and needs to be evaluated (Upmeier zu Belzen & Krüger, 2010). The present article contributes to this issue by analyzing the levels of difficulty and the assumption of an increasing degree of complexity from level I to level III in each aspect.

Forced choice-tasks have been developed to assess students' (*N* = 901) understanding of models and modeling in biology. Two polytomous IRT measurement models have been used to analyze the data: The partial credit model (PCM; Masters, 1982) and the rating scale model (RSM; Andrich, 1978) which both have been developed for tasks with more than two ordered response categories. However the rating scale model assumes equidistant scoring (Andersen, 1977). A comparison of both mod-

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl

els' estimates provides an insight to what extent this assumption is appropriate for the current data (Wu, Adams, & Wilson, 2007) and thus may provide hints concerning specific learning differences in the five aspects.

Theoretical Background

Based on international research findings on models and modeling in science education (e.g. Crawford & Cullin, 2004; Grosslight, Unger, Jay, & Smith, 1991; Justi & Gilbert, 2003) Upmeier zu Belzen and Krüger (2010) developed five aspects concerning the understanding of models and modeling in biology: *Nature of models, multiple models, purpose of models, testing models*, and *changing models*. Following Mahr (2009) three perspectives on models can be distinguished: The model object, the creation of a model, and the application of a model. These perspectives have been used by Upmeier zu Belzen and Krüger (2010) in different ways to develop three levels of understanding in each aspect (Table 1): In the aspect *nature of models* different beliefs about the relation of a model and its original are described. Hence all levels refer to the creation of the model in this aspect. In the aspects *multiple models, testing models*, and *changing models* level I refers to the model object, level II to the creation, and level III to the application of the model. In the aspect *purpose of models* level I and level II refer to the creation of the model but level III to its application. The three levels in each aspect describe an increasing degree of complexity but there is no theoretical description about the difficulties of the adjacent levels in each aspect (Upmeier zu Belzen & Krüger, 2010). So it is not clear if the levels describe equal steps of difficulty across the five aspects or if the step difficulties are specific for each aspect. Terzer and Upmeier zu Belzen (2011) were able to separate the three levels empirically. However a possible hierarchy needs to be evaluated further. For example differences in the levels' relative difficulty between the aspects may provide practitioners with clues about specific learning difficulties in each aspect. In the following it will be analyzed to what extent the levels describe common steps of difficulty across all five aspects and to what extent they describe an increasing degree of difficulty in each aspect from level I to level III. Hence it is assumed that more complex levels of understanding are represented by higher step difficulties (i.e. thresholds).

Table 1. The theoretical structure as proposed by Upmeier zu Belzen and Krüger (2010). Grey indicates the three perspectives by Mahr (2009). Light grey: model object; medium grey: creation; dark grey: application.

	Level 1	Level 2	Level 3
Nature of models	Replication of the original	Idealized representation of the original	Theoretical reconstruction of the original
Multiple models	Different model objects	Different foci on the original	Different hypotheses about the original
Purpose of models	Describing the original	Explaining the original	Predicting something about the original
Testing models	Testing the model object	Parallelize the model and the original	Testing hypotheses about the original
Changing models	Correcting defects of the model object	Revise due to new insights	Revise due to the falsification of hypotheses about the original

Research Questions

(1) To what extent do the data support equidistant differences between the three levels of understanding across all five aspects of models and modeling?

Hypothesis (H1): The differences are not equidistant across all aspects because the theoretical perspectives by Mahr (2009) have not been used consistently by Upmeier zu Belzen and Krüger (2010) to develop the levels of understanding in the five aspects (Table 1). Consequently the PCM is assumed to have a better model fit and smaller information indices than the RSM (Kang, Cohen, & Sung, 2009).

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl

(2) To what extent do the data support an increasing difficulty of the three levels in each aspect?

Hypothesis (H2): Since the three levels are developed as (at least) ordinal categories (Upmeier zu Belzen & Krüger, 2010) the thresholds, the point-biserial correlations, and the person measures in each level are expected to increase monotonically (Linacre, 2002; Wu et al., 2007).

Method

30 forced choice-tasks (6 tasks per aspect) have been developed to assess students' (*N* = 901; 11 to 17 years old; secondary school) understanding of biological models according to the theoretical structure by Upmeier zu Belzen and Krüger (2010). A multi matrix booklet design was conceptualized to keep the number of tasks for every student small (for further information about sample and scoring methods see Krell & Krüger, 2011). ConQuest (Wu et al., 2007) was used to analyze the data using the PCM and the RSM. In these models the probability p of person v answering in response category k on item i is estimated as follows (cf. Nering & Ostini, 2010):

$$P(X_{vi} = k) = \frac{exp\sum_{s=0}^{k}(\theta_v - \delta_{is})}{\sum_{r=0}^{k}[exp\sum_{s=0}^{r}(\theta_v - \delta_{is})]}, \text{ where } \delta_{i0} \equiv 0, \text{ so that } \sum_{s=0}^{0}(\theta_v - \delta_{is}) = 0 \text{ (PCM)},$$

$$p(X_{vi} = k) = \frac{\exp\sum_{s=0}^{k}(\theta_v - (\delta_i + \lambda_s))}{\sum_{r=0}^{k}[\exp\sum_{s=0}^{r}(\theta_v - (\delta_i + \lambda_s))]}, \text{ where } \lambda_0 \equiv 0, \text{ so that } \sum_{s=0}^{0}(\theta_v - (\delta_i + \lambda_s)) = 0 \text{ (RSM)}.$$

In the PCM the thresholds (δ_{is}) are not equidistant within an item and they may vary between items whereas there is only one threshold for all items in the RSM: $\delta_{is} = \delta_i + \lambda_s$, where δ_i is the location of the item on the latent scale and λ_s is the overall threshold. Hence equal distances between the response categories are assumed in the RSM ('equidistant scoring'; Andersen, 1977). The PCM and the RSM are nested models with the RSM as the more restrictive one so they can be compared using descriptive information indices (e.g. cAIC, aBIC) and the chi square difference test (Kang et al., 2009). If the distances between the three levels of understanding are equidistant across all aspects, the RSM will show a better model fit as it is the more parsimonious one. If the assumption of equidistant scoring is not adequately applicable to the actual data the PCM will show a better fit (Wu et al., 2007).

The empirical dimensionality of the construct 'understanding models in biology education' is still an open question (Terzer, Krell, Krüger, & Upmeier zu Belzen, 2011). Since the current analysis focuses on the level of understanding and not on the dimensionality, one-dimensional measurement models are used to estimate the relevant values (cf. research question two). This allows more reliable and more trustworthy estimations (Pietsch, 2010). Using the better fitting measurement model (cf. research question one) the thresholds, the point-biserial correlations, and the average person measures are analyzed. These values should be as follows to support the assumptions of an increasing difficulty: The thresholds should be $\delta_{i1} < \delta_{i2}$, with 1.4 logits < δ_{i2} - δ_{i1} < 5.0 logits (Linacre, 2002) – or λ_s instead of δ_{is} for the RSM. The point-biserial correlations for the three levels of understanding (Wu et al., 2007) and the average person measures (Linacre, 2002) should increase monotonically. To ensure precise estimates Linacre (2002) recommends at least ten observations per category.

Results

The PCM shows smaller values of cAIC and aBIC than the RSM (Table 2). Beyond that the chi square difference test results in a significantly better fit of the PCM: $\Delta\chi^2(\Delta df) = 138.16(29)$;

Table 2. The number of estimated Parameters (nP), the Deviance (Wu et al., 2007), the cAIC (Burnham & Anderson, 2004), the aBIC (Sclove, 1987), and the EAP/PV-reliability (Wu et al., 2007) of the PCM and the RSM.

	nP	Deviance	cAIC	aBIC	Reliability
PCM	61	13,500.99	13,632.01	13,722.28	0.687
RSM	32	13,639.15	13,705.59	13,775.24	0.685

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl

$p < .001$. The reliability (rel.$_{EAP/PV}$ = 0.69) and the variance (var. = 1.00) of the single dimension is acceptable.

For the PCM, MNSQ- and *t*-values indicate a good fit between the data and the one-dimensional PCM (Penfield, 2004; Smith, 2000). These values are comparatively poor for the RSM (Table 3).

Table 3. Unweighted and weighted MNSQ- and *t*-values for the PCM and the RSM.

	Unweighted		**Weighted**	
	MNSQ	*t*	**MNSQ**	*t*
PCM	0.90 ≤ MNSQ ≤ 1.14	-1.30 ≤ *t* ≤ 1.70	0.90 ≤ MNSQ ≤ 1.13	-1.40 ≤ *t* ≤ 1.80
RSM	0.77 ≤ MNSQ ≤ 1.34	-3.00 ≤ *t* ≤ 3.70	0.77 ≤ MNSQ ≤ 1.34	-3.10 ≤ *t* ≤ 3.80

The estimates of the PCM additionally underline the assumption of non-equidistant differences between the three levels across the five aspects. For example there are relatively big distances δ_{i2} - δ_{i1} in the aspect *nature of models* whereas these distances are relatively small in the aspect *purpose of models* (Table 4).

Since the PCM shows a significantly better fit this model is used to evaluate the three levels' difficulty (Table 4). In the aspects *nature of models, multiple models*, and *purpose of models* most of the values are in line with the assumptions of ordered response categories which are outlined above: The differences of the thresholds are between 1.4 and 5.0 logits in most cases and the point-biserial correlations as well as the average person measures increase monotonically from level I to level III. However the difference δ_{I2} - δ_{I1} is greater than 5.0 logits in the aspect *nature of models*. Furthermore the point-biserial correlations of items N_V, M_I, and M_V are quite close but still increase monotonically. In the aspect *testing models* the point-biserial correlations of all items do not increase monotonically. Similarly they do not increase monotonically in the aspect *changing models* in three cases. Finally there are more than ten observations in each response category despite of item N_II (level I) and item T_VI (level I) and the observed counts of level I are relatively small in the aspects *testing models* and *changing models* for all items (Table 4).

Discussion

In line with H1 the PCM shows a better model fit than the RSM. Hence the assumption of equidistant scoring across all five aspects does not seem to be appropriate for the current data. Consequently the relative difficulty of the three levels is likely to vary across the five aspects. For example it seems to be comparatively 'hard' to take the step from level I to level II in the aspect *purpose of models* and comparatively 'easy' to answer on a high level in the aspects *testing models* and *changing models*.

H2 can be confirmed partially since the relevant values indicate an increasing degree of difficulty from level I to level III in the aspects *nature of models, multiple models*, and *purpose of models*. However in the aspects *testing models* and *changing models* the point-biserial correlations indicate that it is 'harder' to answer on level I than on level II. In both aspects level I refers to the model object (Mahr, 2009) and it seems to be hard for students to understand the testing and changing of a model only with respect to the model object. However both the item thresholds and the average person measures are in line with the assumptions of ordinal response categories. Consequently the point-biserial correlations may be disordered because of relatively few students ranking level I first and not because of relatively high cognitive demands of this level (Wu et al., 2007). That is why it should be investigated further if the levels of understanding describe an increasing degree of difficulty, especially in these two aspects.

There are two possible limitations of the present study. First, as mentioned above, the five aspects were treated as one scale (i.e. one-dimensional) in order to allow more reliable estimations (Pietsch, 2010). However, this may have influenced the findings. For example, the RSM's assumption of 'equidistant scoring'

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl

Table 4. The thresholds δ_{i1} and δ_{i2}, the observed count, the point-biserial correlations (bold numbers: $p < .05$), and the average person measures for the items I to VI of the five aspects (Upmeier zu Belzen & Krüger, 2010).

Item	Threshold		Observed count (%)			Point-biserial corr.			Person measure		
	δ_{i1}	δ_{i2}	I	II	III	I	II	III	I	II	III
Nature of models											
N_I	-3.44	1.71	4.20	78.32	17.48	**-0.26**	-0.07	**0.21**	-1.01	-0.16	0.40
N_II	-3.40	1.48	3.92	73.53	22.55	**-0.33**	-0.13	**0.29**	-0.92	0.01	0.56
N_III	-1.72	1.43	16.71	62.40	20.89	**-0.31**	-0.09	**0.39**	-0.56	-0.05	0.64
N_IV	-2.21	1.41	12.80	65.40	21.80	**-0.34**	**-0.19**	**0.49**	-0.89	-0.23	0.87
N_V	-2.47	1.42	9.25	67.62	23.13	**-0.30**	**-0.29**	**0.53**	-0.81	-0.14	0.86
N_VI	-2.53	1.41	8.70	68.90	22.41	**-0.40**	**-0.16**	**0.44**	-1.17	-0.08	0.62
Mean	-2.68	1.48	9.26	69.36	21.38	-0.32	-0.16	0.39	-0.89	-0.11	0.66
Multiple models											
M_I	-2.92	1.18	10.30	62.66	27.04	**-0.37**	**-0.30**	**0.58**	-0.88	-0.08	1.00
M_II	-1.80	0.97	15.01	57.22	27.76	**-0.46**	**-0.11**	**0.49**	-0.90	-0.10	0.56
M_III	-1.85	1.03	14.33	57.67	28.00	**-0.45**	**-0.21**	**0.58**	-0.82	-0.17	0.78
M_IV	-2.06	1.19	12.90	61.29	25.81	**-0.40**	-0.11	**0.42**	-0.84	-0.13	0.62
M_V	-1.92	1.03	12.50	59.44	28.06	**-0.33**	**-0.27**	**0.54**	-0.66	-0.14	0.74
M_VI	-1.76	1.11	15.75	58.27	25.98	**-0.35**	**-0.21**	**0.53**	-0.76	-0.17	0.68
Mean	-2.05	1.09	13.47	59.43	27.11	-0.39	-0.20	0.52	-0.81	-0.13	0.73
Purpose of models											
P_I	-2.53	1.57	8.85	71.24	19.91	**-0.41**	-0.03	**0.33**	-1.09	-0.03	0.59
P_II	-1.97	1.29	14.07	63.47	22.46	**-0.40**	-0.06	**0.40**	-0.84	-0.10	0.52
P_III	-1.67	0.87	16.59	54.15	29.26	**-0.42**	**-0.14**	**0.49**	-0.84	-0.10	0.72
P_IV	-1.48	0.82	18.60	52.98	28.42	**-0.49**	0.02	**0.40**	-0.84	-0.14	0.36
P_V	-1.76	1.04	15.58	57.61	26.81	**-0.41**	**-0.17**	**0.53**	-0.78	-0.18	0.78
P_VI	-1.83	1.18	14.50	60.22	25.28	**-0.37**	**-0.12**	**0.44**	-0.69	-0.06	0.67
Mean	-1.87	1.13	14.70	59.95	25.36	-0.42	-0.08	0.43	-0.85	-0.10	0.61
Testing models											
T_I	-3.16	0.30	4.63	53.70	41.67	**-0.29**	**-0.42**	**0.55**	-1.00	-0.36	0.55
T_II	-2.94	0.25	4.97	50.55	44.48	**-0.31**	**-0.36**	**0.50**	-1.08	-0.26	0.50
T_III	-2.95	0.03	5.17	47.84	46.98	**-0.30**	**-0.38**	**0.51**	-1.17	-0.34	0.37
T_IV	-2.59	0.32	6.69	51.88	41.42	**-0.23**	**-0.45**	**0.57**	-0.65	-0.25	0.40
T_V	-3.04	0.38	5.31	53.75	40.94	**-0.30**	**-0.40**	**0.54**	-1.07	-0.38	0.57
T_VI	-3.75	0.09	2.59	49.74	47.67	**-0.25**	**-0.48**	**0.56**	-1.33	-0.35	0.56
Mean	-3.07	0.23	4.89	51.24	43.86	-0.28	-0.42	0.54	-1.05	-0.32	0.49
Changing models											
C_I	-3.19	0.92	4.63	63.35	32.03	**-0.27**	**-0.30**	**0.44**	-0.96	-0.11	0.56
C_II	-3.21	1.03	4.78	66.89	28.33	**-0.26**	**-0.33**	**0.46**	-0.92	-0.25	0.51
C_III	-2.75	0.65	6.93	58.76	34.31	**-0.40**	**-0.22**	**0.44**	-1.22	-0.25	0.41
C_IV	-2.87	0.71	6.52	59.06	34.42	**-0.40**	**-0.21**	**0.42**	-1.48	-0.20	0.49
C_V	-2.87	1.11	6.25	66.12	27.63	**-0.28**	**-0.34**	**0.51**	-1.05	-0.18	0.77
C_VI	-2.87	0.70	6.81	58.64	34.55	**-0.37**	**-0.30**	**0.50**	-1.17	-0.28	0.61
Mean	-2.96	0.85	5.99	62.14	31.88	-0.33	-0.28	0.46	-1.13	-0.21	0.56

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl

(Andersen, 1977) may be appropriate within each aspect but not, as shown in the present study, across the five aspects. In future investigations separate analyses for each aspect may reveal to what extent there are equidistant thresholds within each aspect. Second, the present findings show and evaluate the relative fit of the RSM and the PCM. However, there are additional IRT models which may also fit the assumptions of the theoretical structure proposed by Upmeier zu Belzen and Krüger (2010) and may be used to gain further insights about students' understanding of models. For instance the ordered partition model (Wilson, 1992) which does not assume an order of all response categories may provide additional information about the relative difficulty of the three levels of understanding models and modeling in biology education.

Acknowledgements

The author wants to thank Dirk Krüger, Eva Terzer and Melanie Jüttner for their comments on earlier drafts of this paper as well as the two anonymous reviewers for their valuable comments and suggestions to improve the quality of the article.

References

Andersen, E. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*, 69–81.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*, 581–594.

Burnham, K. & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*, 261–304.

Crawford, B. & Cullin, M. (2004). Supporting prospective teachers' conceptions of modeling in science. *IJSE, 26*, 1379–1401.

Grosslight, L., Unger, C., Jay, E., & Smith, C. (1991). Understanding models and their use in science: Conceptions of middle and high school students and experts. *JRST, 28*, 799–822.

Justi, R. & Gilbert, J. (2003). Teacher's views on the nature of models. IJSE, 25, 1369–1386.

Kang, T., Cohen, A., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499–518.

Krell, M. & Krüger, D. (2011). Forced Choice-Aufgaben zur Evaluation von Modellkompetenz im Biologieunterricht: Empirische Überprüfung konstrukt- und merkmalsbezogener Teilkompetenzen. *Erkenntnisweg Biologiedidaktik, 10*, 53–68.

Linacre, J. (2002). Understanding Rasch measurement. *Journal of Applied Measurement, 3*, 85–106.

Mahr, B. (2009). Information science and the logic of models. *Software and Systems Modeling, 8*, 365–383.

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

Nering, M. & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Oh, P. & Oh, S. (2011). What teachers of science need to know about models: An overview. *IJSE, 33*, 1109–1130.

Penfield, R. (2004). The impact of model misfit on partial credit model parameter estimates. *Journal of Applied Measurement, 5*, 115–128.

Pietsch, M. (2010). Evaluation von Unterrichtsstandards. *ZfE, 13*, 121–148.

Sclove, S. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.

Smith, R. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199–218.

Terzer, E. & Upmeier zu Belzen, A. (2011, September). *Model competence in biology education - Evaluation of a theoretical structure using multiple-choice items*. Paper presented at the 14th EARLI Conference, Exeter.

Terzer, E., Krell, M., Krüger, D., & Upmeier zu Belzen, A. (2011, September). *Assessment of students' concepts of models and modelling using multiple- and forced-choice items.* Paper presented at the 9th International Conference ESERA, Lyon.

Upmeier zu Belzen, A. & Krüger, D. (2010). Modellkompetenz im Biologieunterricht. *ZfDN, 16*, 41–57.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 4*, 309–325.

Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new paradigm of preference for school science investigations. *Science Education, 92*, 941–967.

Wu, M., Adams, R., & Wilson, M. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Camberwell, Vic: ACER Press.

Published on 01/11/2012 on http://edoc.hu-berlin.de/serl