Is There Need for the 3PL Model? Guess What?

Matthias von Davier

_Educational Testing Service, Princeton, NJ_

Maris and Bechger (2009) talk about the many-to-one relationship between models and probabilities associated with the observed data: the conditional probability of the observed responses may be the same for different sets of model parameters or for different models. The authors have taught us (again) a very important lesson: models are based on assumptions. In particular, latent variable models such as the Rasch model and the 3PL model depend on assumptions that are not easily checked. The authors argue that the same observed data can be fitted equally well by models that look very different from one perspective, but can be shown to be variants of a larger class of models, and are related through a permissible set of transformations of the parameter space.

**WHICH OF THE FOLLOWING STATEMENTS IS TRUE?**

(a) For a person with a hammer every problem looks like a nail.
(b) The appropriate model for multiple choice items is the 3PL model.
(c) The Rasch model never fits real data.
(d) None of the above.

Correspondence should be addressed to Matthias von Davier, ETS, MS 02-T, Princeton, NJ 08541. E-mail: mvondavier@ets.org
I would go for (a). If questioned about their beliefs, psychometricians in one camp would argue
the firm conviction that the Rasch model is mathematically elegant and intuitive as well as plau-
sible for practitioners, pointing out the advantages of a simple model that “counts” every item in
the same way. Psychometricians of another camp would argue that the 3PL is much more flexi-
bile and is suitable to take into account that some item types have a nonzero probability to be
solved by guessing and other random response strategies. This leads us to ask: which of these
models is appropriate for test data of a certain type? Or better: is there a correct
answer to this question?

Unfortunately, choosing between the 3PL and the Rasch model or other variants of item
response theory (IRT) does not become easier even after it is understood that these models are
closely related. If an extraneous principle such as Occam’s razor is used, one may argue in favor of
the simpler model; if the goal is to be more flexible in terms of the ability of the item function to fit
different trace lines, a model with more parameters may seem appropriate. To make matters worse,
there are alternatives that can be substituted for the 3PL model when the issue is to account for ran-
donm response strategies or guessing. These alternatives will be discussed below after introducing a
more general framework useful to position the 3PL among other models for guessing.

**A PECULIARITY OF THE 3PL IRT MODEL**

Guessing versus nonguessing is explicitly distinguished in the 3PL model by the introduction of
the guessing parameter. This indicates that different probabilities of a correct response \(X = 1\)
versus a wrong response \(X = 0\) may exist under these two strategies. In the following, let strat-
egy \(S = g\) denote guessing, whereas strategy \(S = r\) denotes responding based on ability, that is,
not guessing. Consider item \(i\) and person \(v\). Then, the probability of a correct response is

\[
P_{3PL}(X = 1 | i, v) = P(X = 1, S = g | i) + P(X = 1, S = r | v, i),
\]

that is, the sum of probabilities of a correct response while guessing and a correct response while
responding based on ability. Note the 3PL assumes that the first term is independent of the
person’s ability, whereas the second term depends on \(v\). The probability of a correct response in
the 3PL is

\[
P_{3PL}(v, i) = \gamma_i + [1 - \gamma_i] P_{2PL}(v, i),
\]

which is equivalent to

\[
P_{3PL}(v, i) = P_{2PL}(v, i) + [1 - P_{2PL}(v, i)] \gamma_i,
\]

with the probability \(P_{2PL}\) following the 2-parameter logistic model. While the 3PL may have not
been developed with that in mind, the model equation defining the 3PL is similar to expressions
commonly found in multinomial processing tree models (Batchelder & Riefer, 1999) as well as
discrete mixture distribution models (McLachlan & Peel, 2000) and mixture IRT models
(von Davier & Carstensen, 2007). The mathematically identical expressions in (2a) and (2b) imply two very different ways of conceptualizing the process of ability-based responding and guessing:

(a) For each item, the examinee decides randomly whether guessing or ability-based responding is chosen. The probability of guessing is \( P(S = g) = \gamma_i \) and \( P(S = r) = (1-\gamma_i) \). If \( (S = r) \) is chosen, the conditional probability of a correct response depends on the ability of examinee \( v \). However, if guessing is chosen, the correct response will be obtained with certainty.

(b) For each item, the examinee decides whether he/she feels that the correct response was found. This decision is based on the probability of a correct response in the 2PL. If it is decided that the correct response was found, the correct response is indeed chosen with certainty. If the examinee decides that the correct response is not known (with probability of an incorrect response following the 2PL), a response is chosen by guessing.

In (2b), if the respondent believes he knows the correct response he never fails to choose the correct one (the probability of slipping is 0.0). If the respondent does not know the correct response, he is quite aware of this and decides to guess. Even more peculiar is (2a) in that guessing or based-on-ability responding are chosen at random without the involvement of the respondent’s ability. If guessing is chosen, the respondent always guesses correctly, whereas when guessing is not chosen, the correct response is found depending on the respondent’s ability.

Note that we do not know how the response process actually takes place. However, in terms of modeling the marginal probability of a correct response, the 3PL can be written in two mathematically equivalent ways that either corresponds to no slipping when choosing the correct response based on ability, or choosing to guess at random, while never failing when guessing. This peculiarity was noted when developing the necessary specifications and constraints to estimate the 3PL in a generalized latent class framework (Vermunt, 2001) or as a factor mixture model (Asparouhov & Muthen, 2006).

ALTERNATIVES TO THE 3PL IRT MODEL

The HYBRID model (Yamamoto, 1989) is one such alternative approach that explicitly incorporates the idea of two strategies of solving the items in a test: the main ability-based strategy \( (r) \) modeling the response probability using a Rasch model or a 2PL IRT model and the guessing strategy \( (g) \), where the response probabilities are specific to items, but are the same for all respondents choosing this strategy. More specifically, based on equation (1) the hybrid model can be written as

\[
P_{\text{MIX}}(x = 1 \mid i, v) = P(g)P(x = 1 \mid g, i) + P(r)P(x = 1 \mid r, i, v),
\]

which indicates that the probability of guessing the correct response depends on the item, while the likelihood of choosing the guessing strategy is independent of the item. Note that neither the probability of a correct response under the guessing strategy nor under the
ability-based response strategy is 1.0. The probabilities of a correct response under guessing \( P(x = 1|g,i) \) may be constrained or estimated freely. The HYBRID model is a model with a continuous and a discrete latent variable, the discrete variable represents the choice of the strategy (guessing or ability-based responding), whereas the continuous variable represents the IRT-based ability. Extensions of this model include the Grade of Membership (GoM) model (Erosheva, 2003), hierarchical latent class models (Vermunt, 2003), as well as hierarchical diagnostic models (von Davier, 2007). In a recent study, Kubinger and Draxler (2007) have shown that a HYBRID mixture model consisting of a Rasch Model and a guessing class fits data equally well as the special case of the 3PL with all discriminations constrained to be equal. The special case of the 3PL model used by Kubinger and others is exactly the case discussed in Maris and Bechger (this issue). The HYBRID model can be extended to multiple strategy classes and other types of item response data (von Davier, 1995; von Davier & Rost, 1997), as well as to a model for assessing the switch between ability-based responses and random responses due to test speededness (Yamamoto & Everson, 1997).

OTHER CONSIDERATIONS

For a nonmixture alternative to the 3PL model, Bock’s (1972) nominal response model allows the estimation of distractor-level parameters in order to assess—in greater detail than what the 3PL allows—how the choice of correct and incorrect responses depends on one (or more) ability variables. The fact that ability does not play a role in the choice between guessing and responding based on item content in the 3PL has spurred developments that take the ability of a respondent into account also at this level of preresponse decisions. San Martin, del Pino, and DeBoeck (2006) presented a model where the choice between guessing and responding depends on the ability of the respondent. An additional consideration that should play a role in the choice of a model for guessing is the question of how well determined the latent structure representing the variable of interest or the item function really is. Haberman (2005a, 2005b) showed that the 3PL IRT model with unconstrained ability distribution is not identified, and he presented evidence that the 3PL with a normal ability distribution fits data only marginally better compared to a 2PL model. This is yet another indication that the specific assumptions and choices we put into a model cannot tell us necessarily what the “true” item function looks like. What we put into the model will determine our view of the outcomes. The determination of whether a latent trait is assumed to be discrete or continuous, normal or skewed, unidimensional or multidimensional, or a mixture of continuous ability and dichotomous strategy variables should depend on the construct to be measured, rather than on the selection available through one’s favorite software tool or IRT model. Haberman, von Davier, and Lee (2008) have shown that discrete latent variable models, even with small numbers of ability levels per ability dimension, perform quite competitively to normal univariate and multivariate IRT models. In many practical situations the use of a continuous latent ability is by no means justified better by model-data fit than a comparably crude distinction with only a small number of discrete ability levels.

Finally, the question as to whether there is a “correct” model for guessing needs to remain unanswered. While there are a variety of approaches available for modeling guessing,
Unfortunately, practical application of IRT models often picks a model out of tradition rather than out of considerations of how guessing or random response strategies are conceptualized. Maris and Bechger (this issue) show convincingly that the choice of a certain approach by no means yields the only possible explanation or even the best description of the observed data.

REFERENCES


Bock, D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika.* (37)1, 29–51.


