Running head: RULE FOLLOWING AND USE IN THE BALANCE-SCALE TASK

# Rule Following and Rule Use in the Balance-Scale Task

Thomas R. Shultz

Yoshio Takane

Department of Psychology

McGill University

## Abstract

Quinlan et al. (this issue) use Latent Class Analysis (LCA) to criticize a connectionist model of development on the balance-scale task, arguing that LCA shows that this model fails to capture a torque rule and exhibits rules that children do not. In this rejoinder we focus on the latter problem, noting the tendency of LCA to find small, unreliable, and difficult-to-interpret classes. This tendency is documented in network and synthetic simulations and in psychological research, and statistical reasons for finding such unreliable classes are discussed. We recommend that LCA should be used with care, and argue that its small and unreliable classes should be discounted. Further, we note that a preoccupation with diagnosing rules ignores important phenomena that rules do not account for. Finally, we conjecture that simple extensions of the network model should be able to achieve torque-rule performance.

A fundamental debate in cognitive science concerns the best theoretical account of knowledge representation, processing, and acquisition. Two main computational contenders have been the classic symbolic-rule account and the neurally-inspired connectionist account. The classic view is that knowledge is represented in rules whose propositions refer to objects and events, that processing occurs as rules are selected and fired thus generating new symbolic propositions, and that knowledge is acquired by learning these symbolic rules. In many connectionist accounts, active knowledge is represented in fluctuating unit activations and long-term knowledge is represented as connections between units, processing occurs as activations are passed from one layer of units to another, and knowledge acquisition results from adjustment of connection weights. The symbolic view is sometimes referred to as *rule use*, and the connectionist view as *rule following*, to the extent that the environment affords regularities that a neural network can absorb.[1]

The use vs. following debate was joined by Quinlan, van der Maas, Jansen, Booij, and Rendell (this issue) in their critique of cascade-correlation connectionist models of development on the balance-scale task, one of the most frequently modeled tasks in developmental psychology. It is generally beneficial for a computational model to be examined from different perspectives than that of the original modelers. But if problems with a model are found, by either the original or secondary modelers, this need not trigger abandonment of the model. It is often more appropriate to determine whether problems can be fixed, particularly if the model offers useful insights, as cascade-correlation has on several phenomena including conservation (Shultz, 1998, 2006), seriation (Mareschal & Shultz, 1999), transitive inference (Shultz & Vogel, 2004), integration of cues for moving

[1] A neural network functions the same whether the training environment is regular or not. But if the environment is regular enough to be described in rules, then a neural network might learn to behave as if it was following those rules, even though the rules are not explicitly represented as such within the network.

objects (Buckingham & Shultz, 2000), pronoun acquisition (Oshima-Takane, Takane, & Shultz, 1999; Shultz, Buckingham, & Oshima-Takane, 1994), shift learning (Sirois & Shultz, 1998), learning of word stress (Shultz & Gerken, 2005), and habituation of infant attention to auditory (Shultz & Bale, 2001) and visual (Shultz & Cohen, 2004) information, in addition to balance-scale acquisition (Shultz, Mareschal, & Schmidt, 1994).

In this rejoinder, we review available rule-detection methods for the balance scale, document and discuss the tendency of LCA methods to find small unreliable classes, underscore important balance-scale phenomena that rules cannot capture, and speculate about what might be required for neural networks to cover a torque rule.

## Rule Detection for the Balance Scale

Quinlan et al.'s critique relies on their use of a particular method for detecting rules known as Latent Class Analysis (LCA). We begin with a brief comparative review of the principal methods for detecting balance-scale rules. The classic Rule-assessment Method (RAM) examines patterns of performance across six problem types (Siegler & Chen, 2002). Use of rule 1 (weight information) is indicated by a pattern of correct performance on balance, weight, and conflict-weight problems and incorrect performance on distance, conflict-distance, and conflict-balance problems. Rule 2 (weight information, but use of distance when the weights are equal across the two sides) is characterized by the same pattern as rule 1, but with additionally correct performance on distance problems. In rule 3, weight and distance information are both used, yielding correct performance on simple problems, but confusion on conflict problems. Although Siegler suggested that rule-3 users guess on conflict problems, others have emphasized use of other rules, such as addition, in which the side with the larger sum of weight and distance is predicted to

descend (Boom, Hoijtink, & Kunnen, 2001; Ferretti & Butterfield, 1986; Jansen & van der Maas, 1997, 2002; Normandeau, Larivee, Roulin, & Longeot, 1989). Rule 4 is characterized by successful performance on all six problem types. Because these four rules tend to develop in order of their numerical designation, they are often taken as evidence that a child is in a particular stage of development.

To accommodate error variance in human performance, RAM researchers tolerate up to 20% deviant responses from these performance patterns. By this criterion, the cascade-correlation model criticized by Quinlan et al. (this issue) progresses through all four stages when trained on randomly-selected problems with a large bias in favor of equal-distance problems (Shultz, Mareschal et al., 1994).

More recently, several researchers have argued that LCA is a methodologically sounder way to detect rules (Boom et al., 2001; Jansen & van der Maas, 1997, 2002). In exploratory LCA, estimated parameters of a statistically-fitting model differ across latent classes, typically designating homogeneous groups of participants that differ from other groups (McCutcheon, 1987). Individuals can be sorted into the latent classes based on membership probabilities estimated from the model.

Because LCA requires large numbers of participants and does better with small numbers of problems, non-diagnostic problems such as balance and weight problems are often omitted from the test set. Also, in recent research, there is often a systematic attempt to distinguish the addition rule from the torque rule by including among the conflict test problems some that can be solved by either addition or torque and others that can only be solved by torque. Torque is a rotational force applied to a lever, multiplied by its distance from the lever's fulcrum. Whereas the addition rule compares weight and distance sums, the torque rule compares their products.

RAM is favored for its transparency, ease of use with relatively small numbers of participants, convergence with other measures such as verbalization, stability over repeated measurements, prediction of which problems will best promote learning, and consistency across a wide variety of problems including conservation, fullness, shadow projection, and concepts of velocity, time, and distance (Siegler & Chen, 2002). RAM has been criticized for using arbitrary scoring criteria (e.g., 20% tolerance), lack of statistical rigor, and inability to assess rules beyond those emphasized by the theoretical analysis of integrating two dimensions of information (Jansen & van der Maas, 2002). This standard theoretical analysis involves a characterization of rule-based stages (Siegler & Chen, 2002). Children are assumed to start with one dimension, begin to include the other dimension when the first one fails to differentiate cases, eventually use both dimensions but become confused when they conflict, and finally integrate the two dimensions correctly. Although it is unclear how the addition rule could be derived from this stage analysis, it has been noted as a strategy by researchers using RAM (Ferretti, Butterfield, Cahn, & Kerkman, 1985; Normandeau et al., 1989).

LCA is favored for providing a statistical fit between a model and psychological data, avoiding arbitrary scoring criteria, allowing falsification of hypothesized rules, and discovery of new rules (Jansen & van der Maas, 1997, 2002). Siegler and Chen (2002) countered that only the issue of statistical fit uniquely favors LCA because RAM also allows for rule falsification and discovery, and choice of a significance level in LCA is no less arbitrary than a tolerance level in RAM. LCA was further criticized for not providing stable assessments of rule use over short time periods and for requiring several orders of magnitude more subjects than RAM (Siegler & Chen, 2002); it is difficult to find an LCA study of the balance scale with fewer than about 500 participants. These two diagnostic

techniques each have their advantages and disadvantages and it is difficult to decide between them merely by applying them to (usually) different datasets.

## A Problem with LCA: Small, Unreliable Classes

An apparent difficulty with using LCA to diagnose rules and stages, whether in children or in computational models, is its tendency to discover small, unreliable classes that are subject to varying interpretations. For example, independent LCA studies of human balance-scale performance (Boom et al., 2001; Jansen & van der Maas, 1997, 2002; Quinlan et al., this issue) produce small, leftover classes with mutually inconsistent interpretations. Boom et al. (2001) found classes suggestive of Siegler's rules 1, 2, 3, 4, the addition rule, and several infrequent uninterpretable classes. Jansen and van der Maas (1997) found classes for Siegler's rules 1 and 2, the addition rule, and a *no-balance* rule predicting that the scale would not balance, which was said to be difficult to interpret. Jansen and van der Maas (2002) reported classes consistent with Siegler's rules 1-4, addition, a *smallest-distance-down* rule, a *distance-and-guessing-when-weights-are-unequal* rule, a rule that seemed to combine rule 3 with the addition rule, and additional difficult-to-interpret classes. It was the smaller classes that tended to be the most difficult to replicate across these human studies.

The LCAs of computer simulations reported by Quinlan et al. (this issue) likewise discovered small, unreliable classes. Their combined, multi-group LCA obscures differences in results between their two simulations, which might otherwise be regarded as replications despite some procedural differences. In the combined analysis, conditional probabilities (of falling into a particular response pattern given membership in a particular latent class) were restricted to be equal between the two datasets, whereas unconditional probabilities (of being in a particular latent class) were estimated

separately for each of the two simulations. These restrictions on conditional probabilities seem unwarranted given that the two datasets yielded different numbers of classes and different unconditional probabilities for these classes. Restricting conditional probabilities to be identical across the datasets artificially makes them identical even when they are not. With this kind of combined analysis, it is difficult to fully assess reliability of classes between the two simulations. But the fact that the two simulations, when analyzed separately, produced different numbers of classes suggests that some of those classes were not reliable across the simulations.

The point is that LCA seems to regularly produce problematic rule candidates when used in a conventional model-fitting manner. The problem with these extra classes is not difficulty of interpretation; humans and artificial neural networks often produce behavior that is difficult to interpret in terms of rules. The real problem with extra LCA classes is that they are small and inconsistent, suggesting that they might be random and meaningless.

Unless and until LCA techniques are improved, researchers using LCA to detect rules should run enough replications to distinguish systematic performance from random variation. Quinlan et al.'s (this issue) two studies are insufficient for this purpose, not only by being too few but also because replication differences were obscured by artificially restricting conditional probabilities to be identical.

Some of the rules identified in simulations by Quinlan et al. (this issue) have latent class probabilities that are quite small, and these small classes are typically difficult to interpret. Quinlan et al. interpreted these small classes as rule mixtures, side preferences, or balance preferences. For example, their *always-balance* rule had unconditional probabilities of only .03 in the Amsterdam simulation and .00 in the York simulation.

Likewise, their *right-side-bias* rule had unconditional probabilities of .02 in each simulation. Rule mixtures also showed very small unconditional probabilities in one simulation or the other.

In this context, one might invoke Boom et al.'s (2001) distinction between classes and strategies. Classes refer to a set of response patterns that are statistically similar, as revealed by LCA. Strategies (or rules) refer to an interpreted procedure that might generate a statistical class. We agree with Boom et al. that classes that cannot be readily interpreted as being produced by sensible rules should not be treated as rules. They are merely statistical groupings that do not fit a rule interpretation.

The only balance-scale rules to be reliably diagnosed by LCA in humans are rules 1, 2, 4, and addition (Boom et al., 2001; Jansen & van der Maas, 1997, 2002). Ignoring the small and difficult-to-interpret latent classes just noted, it is interesting that Quinlan et al. (this issue) found LCA evidence for rules 1, 2, and addition in cascade-correlation networks. Apart from rule 4, these are the same rules consistently found with children using LCA.[2] Abstracting results across studies thus suggests that problems with constructivist connectionist models of the balance scale are few and might be fixable, if these networks could achieve rule 4, an issue we return to later.

## LCA of Synthetic Data

To see if these trends hold with LCA more generally, we generated synthetic data from ideal addition and torque rules for four hypothetical conflict problems, two of which could be solved by either addition or torque comparisons and two of which could only be solved by comparison of torques. The transition from addition to torque rules is the transition disputed by Quinlan et al. (this issue) for cascade-correlation networks. To

[2] An earlier study (Jansen & van der Maas, 1997) reported that LCA found no human-like rules in a back-propagation connectionist model of the balance scale (McClelland, 1989). In contrast, RAM techniques revealed this model's progression through the first three rules, but no stable rule 4.

allow comparisons with psychological and network data, we generated data for 500 synthetic cases, in three steps. First we put 215 cases in the frequency column of those response patterns representing an addition rule (correct on two addition problems, wrong on two torque problems) and a torque rule (correct on all four problems). Then we distributed the remaining 70 cases randomly, in a uniform distribution, across all 16 response patterns, equivalent to assuming a 3-class LCA model. We replicated this procedure ten times, with different random selections in each replication.

For each of the ten replications, frequencies of response patterns were subjected to exploratory LCA with the LEM program (Vermunt, 1997), using default parameter settings throughout. Model fit was evaluated with the Cressie-Read statistic, a generalization of various chi-square statistics (Cressie & Read, 1984). Following LCA conventions, we started with a 1-class model and incremented classes by 1 until we obtained a nonsignificant Cressie-Read value (indicating that the model fit the data) or ran out of degrees of freedom, whichever came first.

For eight of the ten replications, LCA of frequency data led to rejection of 1- and 2-class models and acceptance of a 3-class model. In the other two replications, a three-class model was also rejected. There were insufficient degrees of freedom to proceed beyond three classes. The conditional probabilities of being correct for the 3-class model are plotted in Figure 1 for replication 5 and in Figure 2 for replication 4. The patterns for torque and addition classes were clear and essentially identical across all ten replications. But patterns for the small, third classes were difficult to label and inconsistent across replications, as portrayed in these Figures. Thus, detection of the torque and addition rules was clear and reliable with LCA, but not so for the small third class. Even though

this third class was generated by random processes, LCA often gave it a sharply non-random appearance through its fitting of conditional probability parameters.

Given that LCA consistently yields unreliable small balance-scale classes in humans, neural networks, and synthetic data, we sought to better understand this tendency by considering the statistical properties of LCA.

## Statistical Sources of LCA Problems

Some of the problems with LCA can be traced to the LCA model and the frequency data used as input. Maximum Likelihood Estimation (MLE) is the common method for parameter estimation in LCA, partly because MLE provides several statistical advantages. However, these advantages obtain only when all of the following conditions are satisfied: 1) the fitted model is correct, 2) the sample size is sufficiently large, and 3) other regularity conditions are met. We discuss each of these conditions in turn and then an epistemological problem.

An LCA model consists of two parts, one statistical and the other parametric. The statistical part assumes independent trials and a multinomial probability distribution of multiple possible response patterns, only one of which occurs in each trial. The parametric part assumes that there are several homogeneous groups in a heterogeneous population, with each group member responding to a set of items independently of other items (the Local Independence assumption – LI). Each group (latent class) is characterized by its size and a set of conditional probabilities of responses to particular items. To satisfy the LI assumption, a large number of latent classes typically must be assumed, but this tends to produce latent classes that are difficult to interpret (Bartholomew, 1987; Hagenaars, 1990; Qu, Tan, & Kutner, 1996).

The benefits of MLE emerge only with a sufficiently large sample. However, what constitutes a large sample is controversial (Hagenaars, 1990; Wickens, 1989). There are $2^n$ possible response patterns when there are $n$ dichotomous items, and reliably estimating the probabilities of these response patterns requires a large number of trials (subjects). There should be at least one, but preferably five or more cases in each response pattern. This condition may be difficult to satisfy in practice, particularly when some response patterns rarely occur, which is sure to happen with exponential increases in number of response patterns. This problem is under active consideration (Bartholomew & Leung, 2002; Hoijtink, 1998; Reiser & Lin, 1999), but there is currently no commonly-accepted solution.

One of the regularity conditions for the standard asymptotic properties of MLE is that LCA parameters must reside in the interior of the parameter space. Ironically however, it is often the case that important parameter values (e.g., crisp rules) are actually on the boundaries of the parameter space with conditional probabilities of 0 or 1, as exemplified throughout LCA research on the balance scale. Although there are some attempts to extend asymptotic theory to cover cases in which estimates are subject to inequality constraints (Dijkstra, 1992; Shapiro, 1985, 1988), the theory then becomes more complicated, and this has not yet been sufficiently digested into the LCA literature or software. The only known practical solutions are resampling methods, such as the parametric bootstrap (Aitkin, Anderson, & Hinde, 1981; Langeheine, Pannekoek, & van de Pol, 1996).

Moreover, even when all three conditions are met, there is an unresolved epistemological issue, namely that there is no statistical method to determine the correct number of latent classes. One may argue that goodness-of-fit tests can determine the

number of significant latent classes. However, such tests are not designed to do this – they are instead designed to determine how many latent classes are needed to satisfy the LI assumption. The number of latent classes naturally increases as the sample size increases, because with a large sample even a small departure of the model from the data becomes significant, and in order to get a satisfactory fit, the number of latent classes has to be increased. With number of latent classes directly dependent on sample size, there is no correct number of latent classes in LCA. The number of latent classes to extract can also depend on the purpose of the analysis. Some researchers want most of the variability in the data explained by a model (say, 99%), whereas others are satisfied with only 80%. The former retain all the latent classes, while the latter keep only the high-frequency classes.

These are the same reasons why the statistical approach to factor analysis, a related technique for finding latent structure, has never found the correct number of common factors defining human intelligence and other characteristics. When the sample size is large, nominally small correlations become significantly different from zero, and a large number of factors are required to explain the correlations. Researchers seeking a simpler, more unified picture of intelligence may opt for smaller samples, whereas those convinced of the complexity of intelligence can support their position with larger samples. Unless and until such statistical problems are solved, our recommendation is to confine interpretation of latent classes to those that replicate across studies.

## A Balance-scale Phenomenon that Rules Do Not Cover: The Torque-difference Effect

Although not as well known as progression through stages, there is another replicated phenomenon in the balance-scale literature known as the torque-difference effect. Torque difference on the balance-scale problem is the absolute difference between the torque on

one side of the scale and that on the other side. The larger this absolute difference, the easier the problem is for children to solve (Ferretti & Butterfield, 1986; Ferretti et al., 1985). This phenomenon has been noted at every stage of balance-scale development in children and in both binary-coded back-propagation (Schmidt & Shultz, 1991) and real-number-coded cascade-correlation (Shultz, Mareschal et al., 1994) network models.

The torque-difference effect occurs naturally in neural networks because they are sensitive to the actual amounts of input signals, and clearer left-right input differences from one side of the scale to the other lead to clearer representations and decisions down stream (Shultz, 2003). This effect is immune to crisp symbolic rules because such rules care only about the direction of differences, not their actual amounts. For example, a symbolic rule characteristic of Stage 1 would specify the side with the larger weight will descend, regardless of how much larger it is. The torque-difference effect is generally considered to be a perceptual effect in that it is based on intuitions about how a balance scale looks on each side. People informally report that they picked one side to descend because it looked like it would. There are many such perceptual effects in the cognitive developmental literature (Shultz, 2003) – they are the rule rather than the exception!

One of the unfortunate aspects of a preoccupation with rule diagnosis is the relative neglect of such perceptual effects. Such neglect is natural for rule-assessment and LCA researchers because crisp rules cannot detect these perceptual effects. This relative neglect encompasses not only an inability to detect perceptual phenomena but also the design of balance-scale problems with relatively restricted ranges of torque difference. In contrast, those willing to look for the torque-difference effect ensure that the training and test sets include problems representative of a wide range of torque differences (Shultz, Mareschal et al., 1994). A more complete evaluation of balance-scale psychology and

models requires coverage of both rule-like consistencies and perceptual effects. One of the apparent virtues of connectionist models has been the integration of such cognitive and perceptual phenomena within a single computational system.

Quinlan et al. (this issue) approvingly cite a symbolic rule-based (ACT-R) model of the balance scale for covering the torque-difference effect (van Rijn, van Someren, & van der Maas, 2003). However, closer examination reveals that this model shows a torque-difference effect only with respect to differences in distance but not differences in weight, and it shows this only in the vicinity of stage transitions, not throughout development as children apparently do. Because of these limitations in this ACT-R model, the torque-difference effect remains uniquely covered by connectionist models.

## Simulating Rule 4 in Balance Scale Development

Another important but contentious main point made by Quinlan et al. (this issue) is that connectionist models of balance-scale development do not capture rule 4 performance, which on some theoretical accounts (Siegler & Chen, 2002) involves computation and comparison of torques. Quinlan et al. (this issue) correctly point out that, because many conflict problems can be solved by adding (rather than multiplying) weight and distance, documentation of a torque rule needs to be supported by success on problems that cannot also be solved by addition. This distinction is a valuable contribution of their and other LCA papers (Boom et al., 2001). It is a distinction that could, and probably should, be added to RAM techniques. Indeed the basic idea of including problems that distinguish between rules is consistent with RAM assumptions.

With five pegs and five weights, the problem size used in cascade-correlation simulations of the balance scale (Shultz, Mareschal et al., 1994), there are 625 total problems, of which just 200 are relatively difficult conflict problems. Only 52 of these

conflict problems (dubbed *torque* problems) actually require a torque rule for correct solution because the other 148 conflict problems (dubbed *addition* problems) can be solved correctly by adding distance and weight on each side and comparing these sums.

Until recently, addition was routinely ignored in both symbolic (Langley, 1987; Newell, 1990; Schmidt & Ling, 1996) and connectionist (McClelland, 1989; Shultz, Mareschal et al., 1994) simulations of balance-scale performance, with no attempt to distinguish addition from a torque rule by including both torque and addition problems in the training and test sets. Thus it is not surprising that LCA methods failed to find evidence of a torque rule, distinct from an addition rule, in replications of older connectionist models (Jansen & van der Maas, 1997).

Quinlan et al. (this issue) conclude that such connectionist models may not be able to learn a true torque rule. This assessment may be too pessimistic because the notion of torque as a product of number of weights and distance from the fulcrum is a rather simple multiplicative function, even if it has to be computed on both sides of the fulcrum and the larger torque selected as marking the descending side of the scale. An obvious way to see if networks can learn to compute and compare torques would be to add more torque problems to the training set. Extensive training for up to 1000 epochs with a highly biased training set, as Quinlan et al. (this issue) tried, does not seem to be an effective way of testing this capacity because the relatively few torque problems that get into the training set are swamped by the much larger number of problems that can correctly be solved by addition or even simpler rules involving weight or distance.

Although we are unable to present new balance-scale simulations in this brief rejoinder, we conjecture that constructive networks could learn a torque rule in either of two ways: by prolonged training with sufficient numbers of torque problems, or by being

taught an explicit torque rule as often happens in secondary-school science courses. The former method could conceivably be implemented in ordinary cascade-correlation networks, the later by newer, knowledge-based technology that permits recruitment of previously learned networks or injected functions as well as single hidden units (Shultz & Rivest, 2001; Thivierge, Dandurand, & Shultz, 2004).

## Rule Following and Rule Use

As noted at the start, an important theoretical distinction is between following and using rules. A symbolic rule-based account of cognition posits that people actually use rules to represent and manipulate knowledge, whereas connectionism assumes that people can behave as though they were following rules because they have learned regularities afforded by the environment. There is also promising work on integrating the two methods, such as with a lower-level connectionist system coupled to a higher level rule system, sometimes also implemented in connectionist fashion (Sun, Slusarz, & Terry, 2005).

Given the relatively new emphasis on distinguishing addition from torque rules at stages 3 and 4, there is not yet a complete connectionist model of balance-scale development. But it may be premature to conclude that such a model is out of the question. It could be a formidable challenge to navigate through all four stages and still terminate with performance that gets nearly all problems correct without using addition, but it is surely worth a try.

## Acknowledgements

This work was supported by operating grants from the Natural Sciences and Engineering Research Council of Canada to each author. We are grateful to Jan Boom,

Frédéric Dandurand, Jay McClelland, Robert Siegler, and Ali Uenlue for helpful suggestions on previous drafts.

## References

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society B, Series A, 144*, 419-461.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. London: Charles Griffin and Company.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse $2^p$ contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*, 1-15.

Boom, J., Hoijtink, H., & Kunnen, S. (2001). Rules in the balance: Classes, strategies, or rules for the balance scale task. *Cognitive Development, 16*, 717-735.

Buckingham, D., & Shultz, T. R. (2000). The developmental course of distance, time, and velocity concepts: A generative connectionist model. *Journal of Cognition and Development, 1*, 305-345.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society B, 46*, 440-464.

Dijkstra, T. K. (1992). On statistical inference with parameter estimates on the boundary of the parameter space. *British Journal of Mathematical and Statistical Psychology, 45*, 289-309.

Ferretti, R. P., & Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Development, 57*, 1419-1428.

Ferretti, R. P., Butterfield, E. C., Cahn, A., & Kerkman, D. (1985). The classification of children's knowledge: Development on the balance-scale and inclined-plane tasks. *Journal of Experimental Child Psychology, 9*, 131-160.

Hagenaars, J. A. (1990). *Categorical longitudinal data*. Newbury Park, CA: Sage.

Hoijtink, H. (1998). Constrained latent class analysis using Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistics Sinica, 8*, 691-711.

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*, 321-357.

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383-416.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrap goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research, 24*, 493-516.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley & R. Neches (Eds.), *Production systems models of learning and development* (pp. 99-161). Cambridge, MA: MIT Press.

Mareschal, D., & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science, 11*, 149-186.

McClelland, J. L. (1989). Parallel distributed processing: Implications for cognition and development. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 8-45). Oxford: Oxford University Press.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Normandeau, S., Larivee, S., Roulin, J. L., & Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *Journal of Genetic Psychology, 150*, 237-250.

Oshima-Takane, Y., Takane, Y., & Shultz, T. R. (1999). The learning of first and second pronouns in English: Network models and analysis. *Journal of Child Language, 26*, 545-575.

Qu, Y., Tan, M., & Kutner, M. H. (1996). Random effects model in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics, 52*, 797-810.

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. In M. Sobel & M. Decker (Eds.), *Sociological methodology* (pp. 81-111). Boston: Blackwell.

Schmidt, W. C., & Ling, C. X. (1996). A decision-tree model of balance scale development. *Machine Learning, 24*, 203-229.

Schmidt, W. C., & Shultz, T. R. (1991). *A replication of McClelland's balance scale model* (No. 91-10-18). Montreal: McGill Cognitive Science Centre, McGill University.

Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika, 72*, 133-144.

Shapiro, A. (1988). Toward a unified theory of inequality constrained testing in multivariate analysis. *International Statistical Review, 56*, 49-62.

Shultz, T. R. (1998). A computational analysis of conservation. *Developmental Science, 1*, 103-126.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI.* (pp. 61-86). Oxford: Oxford University Press.

Shultz, T. R., & Bale, A. C. (2001). Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy, 2*, 501-536.

Shultz, T. R., Buckingham, D., & Oshima-Takane, Y. (1994). A connectionist model of the learning of personal pronouns in English. In S. J. Hanson, T. Petsche, M. Kearns & R. L. Rivest (Eds.), *Computational learning theory and natural learning systems, Vol. 2: Intersection between theory and experiment* (pp. 347-362). Cambridge, MA: MIT Press.

Shultz, T. R., & Cohen, L. B. (2004). Modeling age differences in infant category learning. *Infancy, 5*, 153-171.

Shultz, T. R., & Gerken, L. A. (2005). A model of infant learning of word stress. In *Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society* (pp. 2015-2020). Mahwah, NJ: Erlbaum.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

Shultz, T. R., & Rivest, F. (2001). Knowledge-based cascade-correlation: Using knowledge to speed learning. *Connection Science, 13*, 1-30.

Shultz, T. R., & Vogel, A. (2004). A connectionist model of the development of transitivity. In *Proceedings of the Twenty-sixth Annual Conference of the Cognitive Science Society* (pp. 1243-1248). Mahwah, NJ: Erlbaum. .

Siegler, R. S., & Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology, 81*, 446-457.

Sirois, S., & Shultz, T. R. (1998). Neural network modeling of developmental effects in discrimination shifts. *Journal of Experimental Child Psychology, 71*, 235-274.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review, 112*, 159-192.

Thivierge, J. P., Dandurand, F., & Shultz, T. R. (2004). Transferring domain rules in a constructive network: Introducing RBCC. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (pp. 1403-1409).

van Rijn, H., van Someren, M., & van der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science, 27* 227-257.

Vermunt, J. K. (1997). LEM: A general program for the analysis of categorical data. Tilburg University, Netherlands: Department of Methodology and Statistics.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: Erlbaum.

**figures**

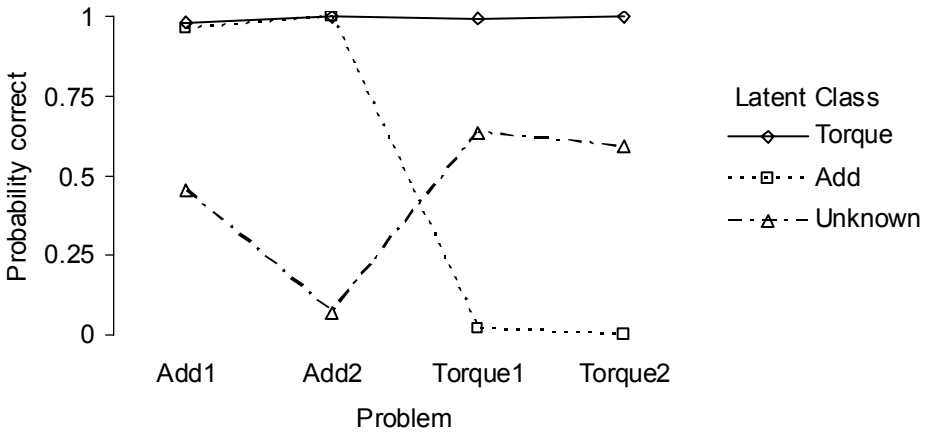


Figure 1. LCA probabilities of being correct on four synthetic balance-scale test problems; replication 5.

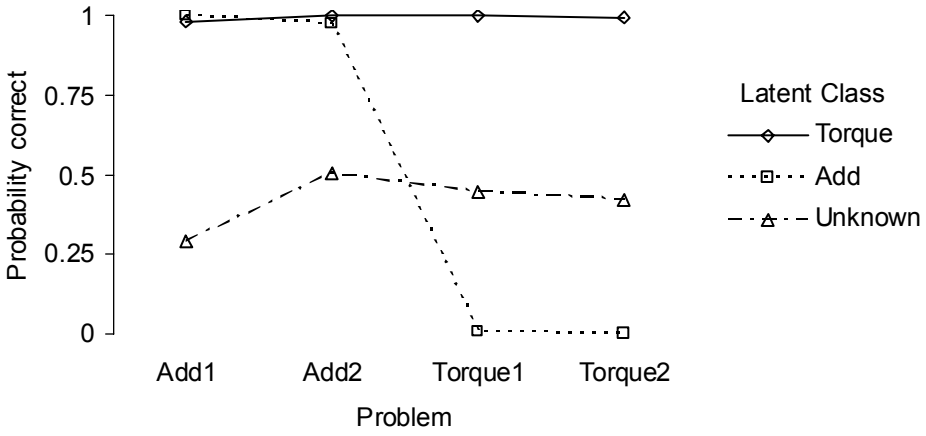


Figure 2. LCA probabilities of being correct on four synthetic balance-scale test problems; replication 4.

Dear Gerry,

Thanks again for inviting us to submit this rejoinder as well as making various suggestions for revision.

We left out the new balance-scale simulations as you requested and clarified the reviewer's issues about conditional and unconditional probabilities.

Best regards,
Tom