Puzzling with potential : dynamic testing of analogical reasoning in children

Stevenson, C.E.

Author: Stevenson, Claire Elisabeth
Title: Puzzling with potential : dynamic testing of analogical reasoning in children
Issue Date: 2012-09-13

CHAPTER 7

General Discussion: Puzzling with potential – the bigger picture

The goal of this thesis project was to develop a new dynamic test of analogical reasoning for school children. The main aims of this thesis were to (1) investigate factors that influence children's differences in performance and change on this new dynamic test of analogical reasoning and (2) examine the predictive value of these dynamic measures on the children's school performance. In this final chapter first an introduction has been provided about AnimaLogica, the dynamic test of analogical reasoning we developed and report on throughout this thesis. In the following two sections investigations from previous chapters into the test design factors and person variables that may affect children's performance and change during dynamic testing have been discussed in reference to the literature. Finally, in section 4, we formulated general conclusions and address theoretical and practical implications.

7.1 AnimaLogica: A dynamic test of analogical reasoning for children

Dynamic testing was introduced in Chapter 1 as a means to measure children's potential for learning in developing cognitive abilities (Sternberg & Grigorenko, 2002). Measuring potential for learning is done by testing and training a child over one or multiple occasions. In AnimaLogica, as with its predecessor the Learning potential of Inductive Reasoning test (LIR, Resing, 1990, the training is provided in the form of graduated prompting techniques (Campione & Brown, 1987; Resing & Elliott, 2011). These interventions are incorporated into the training sessions that are preceded by a pretest and followed by a posttest: i.e., a pretest-training-posttest design. The *pretest* provides an indication of a child's *initial ability* in solving figural analogies (see Figure 7.1) and does not include training or feedback (Resing, 1997). The pretest is a form of static testing and is how conventional tests of cognitive abilities, such as an intelligence test, are usually administered. The pretest is followed by two *training* sessions in which the child receives the graduated prompts training. Graduated prompting involves a standardized protocol of increasingly elaborate instructions starting with metacognitive prompts such as focusing attention, followed by cognitive prompts that explain the solving steps and ending with modeling with scaffolds where the trainer works through the problem step-by-step with the child (e.g., Bosma & Resing, 2012; Resing et al., 2009). An important aspect of the graduated prompts procedure is that instruction is only provided when the child is unable to solve the problem independently, thereby providing information on *instructional-needs*. The *number of prompts* required provides an indication of how much instruction a child needs to reach a particular performance level (Campione & Brown, 1987; Resing, 1993). The *type of prompts* that best aided solution – i.e. metacognitive, cognitive or modeling – may provide information on what type of instruction a child may benefit most from in future interventions (Resing, 2000). The training sessions are followed by a *posttest*, which is tailored instruction – i.e. *potential ability*. The *performance change* in the child's analogy solving from pretest to posttest shows how much can be learned from a short intervention. Examining the child's *self-explanations* and *solution strategies* provides information on the learning process – i.e. how an individual progressed during the dynamic test (e.g., Resing et al., 2009). The ability to solve and explain new but similar *transfer* problems may indicate the depth of learning an individual is capable of after a short, intensive training (e.g., Campione et al., 1985; Ferrara et al., 1986; Resing, 1990).

7.1.1 Main differences with earlier dynamic tests

In AnimaLogica, two problems that have prevented more wide-spread use of dynamic tests were addressed: (1) the extensive duration of administration and (2) the way learning and change is measured (Grigorenko & Sternberg, 1998). The administration of the test developed in this dissertation is considerably shorter than

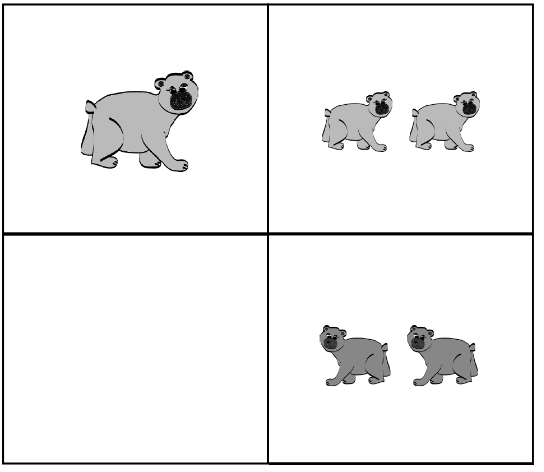

Figure 7.1 *An example figural analogy matrix item.*

previous ones – lasting approximately 80 minutes – and similar to traditional, static cognitive assessment batteries. This efficiency was achieved by providing a shorter training session and limiting assessment of performance on only one task, figural analogies, which could be easily implemented and administered on the computer (see Stevenson et al., 2011 for a discussion of paper versus computer administration). Secondly, the psychometric quality of dynamic tests is often unclear or considered poor as measuring *performance change* is often unreliable from the classical test theory perspective usually used in the statistical analyzes of dynamic tests. The main goal in the (ongoing) development of AnimaLogica was to keep it short and simple, while adhering to rigorous psychometric standards, yet still providing the valuable information unique to dynamic testing about an individual's learning process and cognitive potential.

7.1.2 *Measurement Considerations*

In the dynamic assessment literature, classical test theory measures tend to dominate (e.g., Calero et al., 2011; Resing et al., 2011; Tzuriel & Egozi, 2010). In the typical dynamic testing pretest-training-posttest design, often the posttest percentage correct scores are used as an indication of children's potential ability. However, gain scores (posttest minus pretest score) may be unreliable (Resing, Elliott, & Grigorenko, 2012). Another reason is that change is not necessarily measured on the same scale for test takers with different pretest scores – i.e. it is unlikely that an improvement of 4 correct items is the same if one had 3 or 16 items correct on the pretest. These problems with gains scores could potentially be solved when we use statistical models from item response theory (IRT). In the Rasch model, the simplest IRT model, the chance that an item is solved correctly depends on the difference between the test taker's latent ability and the difficulty of the item. Here the IRT Rasch-based change score has the same meaning across the whole range of the measurement scale in terms of log odds (i.e. the logarithm of probability of correct vs. incorrect), making IRT an appropriate method for measuring change (Embretson & Reise, 2000).

IRT measurement models for dynamic tests have gained some ground (e.g., Hessels & Bosson, 2003; De Beer, 2005). Embretson (1991b) extended the Rasch model and created the Multidimensional Rasch Model for Learning and Change (MRMLC). With this model it is possible to measure initial ability and modifiability (i.e. performance change) from one testing occasion to the next in a dynamic test, without the statistical pitfalls of classical test theory (Embretson, 1987, 1992). In the research with AnimaLogica reported in this thesis IRT models, the MRMLC and the mathematically similar Rasch model for repeated measurements developed by Andersen (Andersen, 1985), were used to measure pretest ability and performance change after training or posttest ability. In Chapter 5 we extended the MRMLC with an explanatory component and thereby demonstrated the usefulness of De Boeck & Wilson's (De Boeck & Wilson, 2004) explanatory IRT approach in a dynamic testing context. Item response theory models hold great promise for dynamic testing and other intervention-based research, not only in reliably measuring differences in individuals' ability to learn, but also in explaining the sources of these differences.

AnimaLogica, as presented in this thesis, uses a non-adaptive item set for the pretest, training and posttest, where the pretest and posttest are isomorphs – i.e. the same problems but with different animals and colors. However, if older children or adults are to be tested then a larger difficulty range is required. In this case computer adaptive testing may be helpful, where the items of appropriate difficulty are selected or constructed from a large pool of possible items during testing (e.g., De Beer, 2005; Embretson, 2004). A downside of computer adaptive testing is that this would require more extensive data collection on item functioning prior to test development than was needed for the fixed item test we created.

A factor that certainly needs to be addressed in future research with AnimaLogica, and perhaps dynamic tests in general, is the scaling of the training items. Item response theory models such as the graded-response model (Samenjima, 1997) or partial credit model (Masters, 1982) seem appropriate for taking the number of required prompts or feedback interventions into consideration when estimating an individual's need for instruction during the training phase of the test (e.g., Attali, 2011; Wang & Heffernan, 2011). Furthermore, the dynamic Rasch model, which assesses whether learning has occurred during testing and the magnitude of individual differences in growth, may be also be appropriate (Verguts & De Boeck, 2000).

7.2 Factors affecting children's performance and change

Children's ability to solve figural analogies develops with great variability throughout childhood evidenced by large differences within each age group both in initial ability as well as performance change (e.g., Siegler & Svetina, 2002; Tunteler et al., 2008). There also appear to be considerable differences between children in the effects of retesting and training of figural analogies (Cheshire et al., 2005; Freund & Holling, 2011b). Similarly, dynamic testing studies show that children generally improve in analogy solving with training, interestingly again with large individual variation in improvement (e.g., Fabio, 2005; Jeltova et al., 2011). Dynamic tests aim to measure individual performance and change in order to gain insight into potential for learning. However, these differences in children's learning during dynamic testing appear to be influenced by test design factors such as training-type or item-format on the one hand and person variables such as working memory or ethnic background on the other hand. AnimaLogica has a number of possible diagnostic outcomes that could be influenced by test design factors: *initial ability, potential ability, performance change, instructional-needs, self-explanations, strategies* and *transfer.* Therefore, the research in this thesis investigated possible factors that could influence the measurement of children's potential for learning with AnimaLogica.

7.2.1 Test design factors

Although numerous aspects of the items or training format may influence children's performance on a dynamic test of analogical reasoning, this thesis was limited to address three of these factors: (1) training-type, (2) test item-format and (3) transfer task choice and administration. How each of these three factors affected children's AnimaLogica performance are now discussed in greater detail.

Much of the focus in the dynamic assessment literature is on "when" the training takes place (while testing or between test sessions) and "how" the training is administered (standardized or not, individually or in a group) (Elliott, 2003; Sternberg & Grigorenko, 2002). AnimaLogica is an individually administered test using a standardized training within a pretest-training-posttest format. This type of dynamic testing format is often validated by comparing a group of trained children with a group that practices independently (e.g., Resing & Roth-Van Der Werf, 2003; Fabio, 2005) or a control condition in which the children receive regular classroom instruction (e.g., L. S. Fuchs et al., 2008) – thus serving as a control for retesting effects or general development. The research reported in this thesis demonstrated that AnimaLogica's graduated prompting training format was generally more effective in improving children's analogy solving than outcome-feedback training (Chapter 5), independent practice (Chapters 3 & 4) or control conditions (Chapters 2 & 3). The graduated prompts training of figural analogies was demonstrated to be an effective means of improving analogy solving with significant large effects comparable to that of other dynamic tests, despite the shorter duration (e.g., Resing, 2000). Furthermore, graduated prompting techniques, which include outcome and strategy-feedback as well as self-explanation prompts, appeared to provide children with more varied learning opportunities which resulted in greater potential results than outcome-feedback, practice or no training. This effect corresponds with work outside of dynamic testing such as the findings of Luwel et al. (Luwel et al., 2010) where strategy-feedback led to greater improvement in children's numerosity judgment than outcome-feedback. In the future, further validation of the strategy-based feedback component within the graduated prompts method could be assessed by comparing it with an outcome-feedback plus self-explanation condition.

A second test design factor investigated in this thesis was the role of item format. This factor has not received much attention in a dynamic testing literature, but as we demonstrated in Chapter 2, may be relevant in gaining insight into a child's potential for learning. Multiple-choice items are often used in cognitive ability assessment, yet this may not be appropriate for dynamic testing as we were more interested in the problem solving process and not just if the child can select the correct answer option. We examined whether training during dynamic testing with multiple-choice or constructed-response items led to differences in children's analogy solving with regard to *strategy progression*, *self-explanation* or *performance change* from pretest to posttest. One group of children was trained on multiple-choice items (MC) and the second group was trained using constructed-response (CR) items – here they had to "construct" the answer in the empty box using a set of animal figures. The results did not show differences in *performance change* from pretest to posttest. The *number of prompts* the CR-trained children required was greater than that of the children in the MC-group, indicating that the CR-items were generally more difficult. Yet, children trained with CR-items provided better quality *self-explanations* compared to those trained with MC items. Also, a difference in *strategy progression* during training between the two training groups was apparent. Duplication is a commonly used strategy by young children who do not yet understand analogical reasoning; it refers to the answer being a copy of the figure in the adjoining box. This non-analogical strategy was used more often by the MC-group whereas the CR-group used a more advanced analogical strategy, partial correct. CR-items appeared to positively affect the children's understanding of analogical reasoning evidenced by better self-explanations and more advanced strategies, despite the greater difficulty of the items. This result coincided with other research in which more active processing has a greater learning effect (Harpaz-Itay et al., 2006; Martinez, 1999). Furthermore, CR-items provided more fine-grained analysis of the children's strategy-use and would therefore simplify diagnosis of erroneous reasoning (e.g., Birenbaum & Tatsuoka, 1987; Birenbaum et al., 1992). CR-items may be very beneficial for process-oriented diagnostics, with the goal of adapting instruction to individual needs where the analysis of strategy progression and extent of understanding are of particular interest (e.g., Grigorenko, 2009a; Jeltova et al., 2007).

The third test design factor addressed in this thesis (Chapter 4) was which task can best be used to measure transfer and when to time the administration. Here we found that performance on the figural analogies pretest was strongly related to performance on the three possible transfer tasks we investigated: geometric analogies, seriation and analogy construction (e.g., Carpenter et al., 1990; Roth-Van Der Werf et al., 2002; Sternberg & Gardner, 1983). Yet, as with previous research on graduated prompting and transfer (e.g., Tunteler & Resing, 2010) we did not find differences in transfer between children trained with graduated prompts or those who practiced independently. Furthermore, children in both groups showed little improvement on the geometric analogies and seriation transfer tasks that were administered during the pretest and posttest sessions. Transfer is notoriously difficult to elicit in experimental settings (Barnett & Ceci, 2002). A possible explanation for our results and those of previous studies where training on a different task does not affect transfer of knowledge to similar tasks stems from Opfer and Thompson's (Opfer & Thompson, 2008) practice interference hypothesis. Their theory suggests that practice using incorrect solution strategies, which often occurs during pretesting, impedes transfer. This hypothesis was supported by the fact that transfer of analogical reasoning skills was only found to the reversal task, in which the child constructed an analogy for the examiner, which was not pretested. Reversal performance was related to *initial ability* on the figural analogies tasks, where more complex analogies were constructed by the children with higher pretest scores. The findings on the reversal task were in line with Siegler's theory (2006) that greater mastery of task strategies increases the chances of knowledge transfer to a novel situation in children. In the assessment of transfer within dynamic tests, which often comprise a pretest-training-posttest format, it is perhaps advisable not to pretest the transfer tasks. Instead a selection of transfer tasks that measure similar skills to the tested task may provide more reliable measures. The effect of initial ability could be accounted for using the pretest scores of the trained task, which indeed correlated with performance on the analogy construction (reversal) task in the present study.

7.2.2 *Person variables*

Different dynamic tests have been developed with different populations in mind, from typically developing children (e.g., Resing & Elliott, 2011), intellectually or developmentally disabled persons (e.g., Hessels, 2009; Hessels-Schlatter, 2002) to clinical populations (e.g., Wiedl, Schöttke, Green, & Nuechterlein, 2004). AnimaLogica focuses on both typically developing elementary school children as reported in this thesis or those in a clinical educational setting (e.g., Resing, Bosma, & Stevenson, 2012). In both of these populations three so-called person variables are often reported in the literature that appear to influence children's performance and change on figural analogies: (1) cultural background, (2) working memory and (3) initial ability. The roles of each of these three factors in children's AnimaLogica performance were investigated in this thesis and are now discussed in greater detail.

Cultural background appears to play a role in performance on cognitive ability measures (e.g., Sternberg et al., 2007). For example, persons from the dominant culture generally obtain higher scores on measures of intelligence or reasoning ability (e.g., Te Nijenhuis & Van Der Vlier, 2001; Van de Vijver, 2002, 2008). These differences in conventional measures can be due to cultural bias in the tests themselves (i.e. item bias), the testing situation (e.g., nonnative instruction language, cultural influences on test-wiseness) or cultural differences in the tested construct (Grigorenko, 2009b; Van de Vijver & Poortinga, 1997). Dynamic testing appears particularly valuable in groups that may be at a cultural disadvantage with traditional testing situations, such as ethnic minority populations, as the training opportunities can perhaps compensate for differences in test-wiseness or non-native instruction language (e.g., Hessels, 2000; Lidz & Pena, 1996; Tzuriel & Kaufman, 1999). Given our aim of developing a dynamic test that could easily be used in diagnostic practice it seemed imperative to consider the culturally diverse backgrounds of many school children in the Netherlands. Figural analogies were chosen as these are considered relatively culture-fair (Cattell, 1979). However, even such items may still be culturally biased (e.g., Van de Vijver, 2002). In Chapter 3 we examined the applicability of AnimaLogica in the dynamic testing of culturally diverse school populations in the Netherlands. In this study, the performance of 7-8 year old children with Dutch parents were compared to that of children with one or both parents from a different country (i.e. ethnic minority children). After confirming that the AnimaLogica items were not biased for one of the two groups, we investigated whether there were differences in their analogy solving progression during dynamic testing. Ethnicity was found to be related to *initial performance* on AnimaLogica as indigenous Dutch children obtained on average higher ability estimates on the pretest than ethnic minorities (e.g., Hamers et al., 1996; Tzuriel & Kaufman, 1999; Van de Vijver, 2002, 2008). However, no differences in *performance change* were found between indigenous and ethnic minority children. This result coincides with previous investigations into cultural differences on dynamic tests (Hamers et al., 1996; Tzuriel & Kaufman, 1999; Sternberg et al., 2007; Resing et al., 2009). Furthermore, we found that *instructional-needs* did not differ as both the number and type of required prompts during training were similar between the two groups. Also, the *self-explanations* of the indigenous Dutch and ethnic minority children did not differ. Cultural bias may still be present when ability is interpreted in the traditional sense as ethnic minorities have systematically lower pre-test scores (Van de Vijver, 2008). However, dynamic measures, quantified by *performance change, self-explanations* and *instructional-needs*, did not appear to suffer from this bias. Dynamic testing may therefore potentially play a more prominent role in the culture-fair assessment of multicultural groups (Grigorenko, 2009b). Future investigations of AnimaLogica as an instrument for multicultural assessment should examine topics of cultural bias and equivalence in more depth.

A second factor that was investigated was working memory, which was addressed from different perspectives in Chapters 3, 4 and 5. Working memory refers to the ability to hold and manipulate entities in memory and shows large increases in childhood (e.g., Swanson, 2008). The role of age in analogy solving has been addressed in the earlier literature. Older children generally perform better on tests of analogical reasoning than younger children (e.g., Siegler & Svetina, 2002; Sternberg & Rifkin, 1979). In Chapter 5 we demonstrated that age is related to *initial ability* on the figural analogy problems, however this relation was confounded by working memory capacity. Research has linked children's performance on fluid reasoning tasks, such as figural matrices, to their memory span and working memory capacity (e.g., Hornung et al., 2011; Kail, 2007; Tillman et al., 2008). We found that working memory capacity (WMC) was related to *initial ability* on AnimaLogica, whereby children with greater working memory had higher ability estimates. Yet, children with greater working memory efficiency did not profit more from graduated prompting than those with smaller working memory capacity – in other words working memory was unrelated to *performance change* in each of these studies. These results corroborate with those of Resing, Xenidou-Dervou, Steijn and Elliott (2012) in which the children also received graduated prompting on a different inductive reasoning task. The graduated prompts procedure provides step-by-step cognitive prompts of how to solve the tasks by attending to each transformation separately. A possible explanation for our findings is that this sequential approach teaches the children a strategy to reduce the cognitive load of the task and thereby improve performance beyond that of control groups regardless of working memory efficiency. This idea is supported by our finding that lower WMC children required more cognitive prompts during the training yet improved their analogical reasoning to a similar extent as the children with higher WMC. A second possibility is that the graduated prompting procedure offers problem solving strategies or feedback that aids the children in more efficient use of their available working memory capacity. This is possibility seems supported by the results of Mackey, Stone, Hill and Bunge (2010) who found that performance on working memory tasks increased with an eight-week figural analogy training. However, in this case it concerns more intense training, therefore in future research WMC measures should be included both before and after training and help determine whether WM efficiency is affected by the graduated prompts intervention.

The third person variable that appears to play a role children's performance on a dynamic test is their initial ability – i.e. what they already know about solving analogies prior to training. We found that children with lower pretest scores generally improved more after the graduated prompts training than children with high *initial ability*, which given the moderate difficulty of the test items and the use of IRT estimations could not be due to ceiling effects (see Chapters 3 & 5). Our finding is in line with those of Swanson and Lussier's meta-analysis of dynamic testing effects who concluded that children with initially lower cognitive ability scores tend to improve more during short dynamic testing training-phases (Swanson & Lussier, 2001). Furthermore, in training studies outside of the dynamic testing domain similar results are found. In the case of Luwel et al. (2010) children with lower intelligence test scores improved more with strategy-feedback compared to children with high intelligence scores. Also, Jaeggi et al. (2008) found that low ability children tended to improve more so than high ability children on figural matrices after training on a working memory task. This finding indicates that children with untapped potential for learning are more often present in groups of low functioning children, but would perhaps be overlooked if they were judged based on a conventional reasoning test. It also appears that the IRT-based measure of *performance change* is more suitable in identifying these children than measures of *instructional-needs* as the number of required prompts in training correlates more strongly with initial ability than with performance change (see Chapter 6).

7.3 Predictive value

The final puzzle piece we investigated was whether recent school performance was related to analogy solving and improvement during dynamic testing. The main aim of Chapter 6 was to investigate predictive value of dynamic testing outcomes on young children's school achievement in reading and math. Dynamic measures may provide additional predictive value of school achievement in reading (e.g., Bynre et al., 2000; D. F. Fuchs et al., 2011; Swanson, 2011b) and math (e.g., Beckmann, 2006; Jeltova et al., 2011; Resing, 1993; Sittner Bridges & Catts, 2011). However, dynamic testing studies do not consistently show advantages of dynamic measures in predicting achievement (e.g., Caffrey et al., 2008). Furthermore, a variety of dynamic measures have been used to predict achievement and it is unclear which dynamic measure (e.g., *potential ability, performance change, instructional-needs, transfer*) is most useful. We compared the predictive value of AnimaLogica'static measure, the pretest score, to three dynamic measures: performance change, instructional-needs and transfer score. The static measure, i.e. the figural analogies pretest, was strongly associated with math achievement, but was surpassed as a correlate of achievement by *instructional-needs* – i.e. the amount of instruction the child needed to correctly solve the training items. In Chapter 3 we had already seen that the children's instructional-needs correlated strongly with teacher ratings and learning ability – which may mean we are tapping into similar information the teacher obtains in the classroom on individual children's ability to learn from instruction (Bosma & Resing, 2012). Yet, instructional-needs and the children's transfer score from the reversal task were often more strongly related to academic achievement measured in the same time period, but not necessarily to subsequent achievement. The dynamic measure of *performance change* provided additional predictive value of reading and math achievement over the course of three measures within one year. This result coincides with Freund and Holling's (2011b) finding that children with higher school grades show the greatest improvement upon retesting. Furthermore, our findings were in line with previous research on the predictive validity of dynamic testing, where performance change, the posttest and/or training scores are better or additional predictors of statically administered measures (Beckmann, 2006; Jeltova et al., 2011; Resing, 1993). The unique contribution of this study was the longitudinal design in which future rather than concurrent achievement was predicted and the identification of which of the dynamic measures provide the best prediction.

7.4 Conclusion

On the whole, children showed great variation in their potential for learning to solve analogies. As with previous research on children's analogy solving progression, the children's performance generally improved over repeated testing occasions, but the degree of improvement varied greatly (e.g., Freund & Holling, 2011b; Siegler & Svetina, 2002; Tunteler & Resing, 2007c, 2007b). The large individual differences in performance and change after the short dynamic testing intervention coincides with findings in other cognitive tasks such as visuospatial reasoning (Embretson, 1987), series completion (Resing, Xenidou-Dervou, et al., 2012) and numerical estimation (Siegler, 2006; Luwel et al., 2010). In AnimaLogica this variation was present in each of the investigated dynamic measures: *strategy-progression, self-explanations, performance change, instructional-needs* and *transfer.* The type of training influenced each of these measures of AnimaLogica performance (Chapters 2 - 5). Also, the item format affected *performance change, strategy-progression, self-explanations* and *instructional-needs* (Chapter 2). *Transfer* performance was related to initial ability and working memory (Chapter 4). Yet, the person variables we investigated, ethnicity and working memory, were not related to *performance change* (Chapters 3, 4 & 5). With regard to ethnicity this technically negative finding is in fact positive as similar dynamic outcomes (*performance change, self-explanations* and *instructional-needs*) between indigenous Dutch and ethnic minority children seems to indicate that AnimaLogica may be an appropriate measure for culturally diverse school children (Chapter 3).

However, given the importance placed upon working memory in cognitive and psychoeducational assessment (e.g., Pickering & Gathercole, 2004) it was important to investigate whether working memory could explain children's differences in the *performance change* and *transfer* on our dynamic test. We found working memory was unrelated to both aspects. Performance change and ability to transfer knowledge to novel situations, such as in the reversal task, are not often included in the assessment of intellectual abilities (Bosma & Resing, 2006; Elliott et al., 2010), yet the findings in this thesis indicate that these two dynamic measures may be separate constructs and important in the assessment of learning and cognitive potential.

Initial ability does seem to affect how children progress in analogy solving during dynamic testing. For example, higher ability children generally require fewer prompts (Chapters 3 & 6) and show greater *transfer* on the reversal task (Chapter 4). Yet, lower ability children tended to show greater *performance change* (Chapter 5). This finding is important because it demonstrates that the children with untapped potential are most likely to be found at the lower end of the spectrum of static testing scores (e.g., Swanson & Lussier, 2001).

The predictive power of our dynamic measures of analogical reasoning – especially Rasch-scaled performance change – above that of static measures confirmed our hypothesis and adds to the growing evidence of the predictive value of dynamic testing in psycho-educational assessment (Chapter 6). Analogical reasoning is often measured in cognitive ability tests (Freund & Holling, 2011a) and has been demonstrated to predict math and reading achievement (e.g., Balboni et al., 2010). Our finding that the dynamic measure of performance change is only somewhat related to *initial ability* and appears to be a better predictor of math and reading achievement, provides further evidence that this may be a separate construct important in the assessment of learning and cognitive potential. Furthermore, the performance change measure, which has often been criticized as a measure of learning potential in the context of classical test theory (e.g., Sternberg & Grigorenko, 2002), has demonstrated its worth when estimated using item response theory models and will hopefully find its place again among the valuable measurement outcomes of potential for learning.

AnimaLogica outcomes appear to be a valuable addition to conventional tests in the prediction of scholastic achievement and applicable for culturally diverse school populations. Furthermore, process-oriented diagnostic information, such as *performance change*, *instructional-needs*, *self-explanations*, *strategies* and *transfer* are available. This information may prove useful for educators in providing interventions that help children more thoroughly utilize their potential for learning at school (e.g., Bosma & Resing, 2012; Jeltova et al., 2011).