Running Head: CLOSER TO DEDUCTION IN TRANSITIVE INFERENCE

Getting One Step Closer to Deduction: Introducing an Alternative Paradigm for Transitive Inference

Barlow C. Wright

Brunel University, School of Social Sciences, Middlesex, England.

Donna Howells

Brunel University, School of Social Sciences, Middlesex, England.

All correspondence and requests for reprints should be addressed to:-

B. C. Wright, Brunel University, School of Social Sciences, Uxbridge, Middlesex. UB8 6RB.

England. Email: barlow.wright@brunel.ac.uk.

Abstract

Transitive inference is claimed to be “deductive”. Yet every group/species ever reported apparently uses it. We asked 58 adults to solve 5-term transitive tasks, requiring neither training nor premise learning. A computer-based procedure ensured all premises were continually visible. response-accuracy and RT (non-discriminative nRT) were measured as typically done. We also measured RT confined to correct responses (cRT). Overall, very few typical transitive phenomena emerged. The symbolic distance effect never extended to premise recall and was not at all evident for nRT; suggesting the use of non-deductive end-anchor strategies. For overall performance and particularly the critical B?D inference, our findings indicate deductive transitive inference is far more intellectually challenging than previously thought. Contrasts of our present findings against previous findings, suggests at least two distinct transitive inference modes, with most research and most computational models to date targeting an associative mode rather than their desired deductive mode. This conclusion fits well with the growing number of theories embracing a “Dual Process” conception of reasoning. Finally, our differing findings for nRT versus cRT, suggests researchers should give closer consideration to matching the RT measure they use to the particular conception of transitive inference they pre-held.

Key Words: Adult Reasoning, Dual Process Theory, Relational Reasoning, Symbolic Distance Effect, Transitive Inference.

Getting One Step Closer to Deduction: Introducing an Alternative Paradigm for Transitive Inference

When a reasoner coordinates two or more pieces of information to deduce a new conclusion, and those pieces of information overlap in a linear way, then s/he may have engaged in Transitive Inference (Goel, Makale & Grafman, 2004; Goodwin & Johnson-Laird, 2005; Martin & Alsop, 2004). Consider the following example. During your lunch break yesterday, you might have gone out running with one friend (say Jane), and noted that you really struggled to keep up with her. We can term the above information a premise, where A and B are entities and there is some comparison made between them (i.e. premise A:B). Now, during today's lunch break you ran the same route with Kerry, and noted she was really struggling to keep up with you (premise B:C). At the end of the day, Kerry tells you Jane has asked her to go running during tomorrow's lunch break, asking you whether you think she will struggle or not (A?C). A transitive inference allows you to work out that, exceptional circumstances aside, Kerry will struggle running with Jane. The beauty is that you can infer such an outcome, without having already experienced it directly. Many argue that the capacity for transitive inference is logical (i.e., deductive), partly because the conclusion necessarily follows as long as the relation used is a linear comparative term (e.g., "runs faster than"), and partly because we can deduce the conclusion upon simply being told the two premises, rather than having to directly perceive and verify them as true for ourselves.

Transitive inference may lay at the heart of a plethora of cognitive and sub-cognitive competencies, from spatial navigation through to predicting where to find food; and from placing oneself within a social network through to scientific thinking (Allen, 2006; Archie et al., 2006; Bond, Kamil & Bolda, 2003; Hummel & Holyoak, 2001; Markovits & Dumas, 1999; Siemann & Delius, 1998; Wright, 1998a, 2001). Further, transitive tasks have been used as a tool in better understanding the similarities and differences between the mental processes of humans and non-humans (Eichenbaum, 2001; Lazareva & Wasserman, 2006; Wu & Levy, 2001). It is for these

reasons among others, that some theorists argue that transitive inference potentially is an important window on cognition (Allen, 2006; Goel et al., 2004).

The above potential remains largely unfulfilled to date. This is because of a number of protracted and diversionary debates about who can solve transitive tasks, at what age, and by what means (Brainerd & Reyna, 1992; Bryant, 1998; Bryson & Leong, 2007; Holcomb, Stromer & Mackay, 1997; Lazareva & Wasserman, 2006; Markovits & Dumas, 1992; McGonigle & Chalmers, 1992; Moses, Villate & Ryan, 2006; Russell, McCormack, Robinson & Lillis, 1996; Shafir, Waite & Smith, 2002; Siemann & Delius, 1996, 1998; Yamazaki, 2004). The debates flourish because of the pivotal role transitive inference was afforded in Piagetian theory and the resultant close scrutiny of both (Bouwmeester, Vermunt & Sijtsma, 2007; Breslow, 1981; Chapman, 1999; Piaget, 1970). The Piagetian classical 3-term task had loosely followed recommendations from cognitive theorists working in the area of reasoning or logic (Bara, Bucciarelli & Lombardo, 2001; Demarais & Cohen, 1998; Goel et al., 2004; Goodwin & Johnson-Laird, 2005). Thus, the minimum two premise pairs were used (A:B and B:C), with the reasoner typically required to make the inference between items A and C (Hong & Chond, 2001; Sternberg, 1980; Wright & Dowker, 2002).

However, in a seminal paper, Bryant and Trabasso (1971) convincingly argued that the Piagetian 3-term task on the one hand could be passed without having to make an inference at all, but on the other hand may be failed because the premises may not be in memory at the time the inference is requested. These false-positive and false-negative arguments are extensively discussed elsewhere (e.g., Brainerd & Reyna, 1992; Wright, 2001). Here, it is sufficient to note that Bryant and Trabasso circumvented them by increasing the number of premises to four (A:B, B:C, C:D and D:E) and repeatedly training participants on these premises. The new task was called the IP-task which is short for the Information Processing task. It introduced to transitive research constructs like cues, memory encoding, symbolic mental representation, and response-time (RT), which had recently been brought together as part of the generic information processing

approach to cognition (Trabasso, 1977). However, giving Bryant and Trabasso's information processing transitive task the same label as an entire perspective used in psychology is potentially confusing. To avoid such confusion, we follow Russell et al.'s (1996) convention of referring to Bryant and Trabasso's task as the B&T task.

The initial B&T finding was that transitive inferences are made by 4 years, which is around half the age estimate from the original Piagetian task (see also Holcomb et al., 1997). Then, since Bryant and Trabasso's original demonstrations, over 90% of transitive studies have followed the B&T methodology of training participants on at least four premise pairs, for as long as it takes to reach near perfect performance (e.g., Acuna, Sanes & Donoghue, 2002; Holcomb et al., 1997; Lazareva & Wasserman, 2006; Martin & Alsop, 2004; Wright, 2006b). Of great significance, in addition to unexpectedly high performance in other human groups (Maydak, Stromer, Mackay & Stoddard, 1995; Stromer, Mackay, Cohen & Stoddard, 1993), almost any non-human group tested has passed the B&T task: From as large as the beluga whale or the elephant (Archie et al., 2006; Murayama & Tobayama, 1997) to as small as the jay or honey bee (Bond et al., 2003; Shafir et al., 2002).

Despite highly contrasting findings between the B&T task and its 3-term predecessor (e.g., on age of reaching competence), many insist B&T tasks target precisely the same "logical" competence as the Piagetian task (Acuna et al., 2002; Bouwmeester et al., 2007; Bryant, 1998; Halford & Andrews, 2004; Yamazaki, 2004). Intriguingly, no theorist seems yet to have offered any rationale for exactly why the B&T task which requires 10 premises to be stored in memory, involves five interlinked items, and tests for no less than six inferences, should either be equivalent to or easier to solve than the 3-term task requiring only one inference to be made. Indeed, we may add Wright's (2001) contention that the reasoner may consider some premises to be reversed in order and alternated in markedness (e.g., may perceive E<D instead of D>E). Additionally, as Goodwin and Johnson-Laird (2005) point out, the reasoner may interpret some premises as using negation of relations (e.g., E not as big as D, instead of E smaller than D).

Adding in these considerations ought to lead to the conclusion that the total number of combinations renders the B&T task extremely difficult compared to its 3-term counterpart, and perhaps particularly for children (Wright, 1998b, 2006a). In support of Wright's postulate, Viskontas, Holyoak and Knowlton (2005) found that even when a 5-term task avoids having to memorise any information, such a task is very much more difficult than 3-term problems.

Over 3,000 research articles have now been devoted to transitive inference. The current state of the debate has led Bryant, arguably the most important contemporary figure in this area, to conclude that "The question of children's ability to make transitive inferences is one of the most vexed in the field of cognitive development." (Bryant, 1998, pp.266). This view is echoed in some theoretical treatments of transitivity (e.g., Allen, 2006; Markovits & Dumas, 1992; Russell et al., 1996). The overarching goal of the present paper was therefore to help reach an understanding of transitive inference that can integrate across perspectives. There were three more specific aims. The first was to determine whether removing the memory demands of a 5-term task would result in it becoming easier than its B&T task equivalent (e.g., Wright, 1998b, 2006a). The second aim was to determine whether response profiles almost invariably accepted in this area are as universal as most believe (Wright, 2006b). Today, most of the acclaimed B&T task profiles actually relate more to RT than to response-accuracy (e.g., Breslow, 1981; Holcomb et al., 1997; Hummel & Holyoak, 2001). The final aim was therefore to provide important data on the appropriateness of two different measures of RT. We expand on these aims below.

1, Towards a Valid Normative Model of Transitive Inference

The first aim reduces to establishment of a limiting adult model – i.e., what level of competence is to be regarded as typical from early adulthood onwards? Such a model is essential if we are to understand transitive inference in children and non-humans. Breslow (1981) was among the first to address this question (see also, Goodwin & Johnson-Laird, 2005; Trabasso, 1977). However, all the data being modelled were tied to the B&T task. Markovits and Dumas (1992) and Russell et al. (1996) claimed that some B&T tasks (namely those developed for non-humans)

unintentionally assess a non-deductive transitive competence. Wright (2001) claims that this argument might also apply to different transitive tasks used with humans. In response, some theorists argue for the existence of two or more processes of transitive inference (Allen, 2006; Bryson & Leong, 2007; Goel et al., 2004; Lazareva & Wasserman, 2006; Markovits & Dumas, 1999; Reyna, 2004; Schnall & Gattis, 1998; cf. Wright, 1998a). Such a postulate is directly in line with recent thinking about modes of reasoning more generally. Most notably, Evans (2003) discusses a Dual Process account of reasoning. This theory argues for the existence of an evolutionary-older network of brain structures for "associative" reasoning, plus a more specialist newer network for "deductive" reasoning (see also De Neys & Glumicic, in press; Ferreira, Garcia-Marques, Sherman & Sherman, 2006; Kokis et al., 2002; Reyna, 2004).

In contrast to dual process accounts of reasoning and of transitive inference in particular, some researchers continue to hold that transitive tasks may differ but transitive inference is one and only one cognitive ability (Bryant, 1998; Halford & Andrews, 2004; Wynne, 1998). Regardless of which conclusion one currently prefers, it remains the case that limiting research largely to the B&T task might restrict or even cause the types of theory that might be developed about other aspects of transitive inference (Wright, 1998a). For example, Brainerd and colleagues have conducted considerable research into the question of whether inferential responses derive from "gist" or from the premise information stored verbatim in memory (Brainerd & Kingma, 1984; Brainerd & Reyna, 1992). However, it may be that during day-to-day deductive transitive inference, reasoning about more than three-relations plus having to memorise all the premise information might be more of an exception rather than the rule (Allen, 2006; Goodwin & Johnson-Laird, 2005). Indeed, this may partly underlie the recent revision of gist theory, which was originally applied to transitive inference (Brainerd & Kingma, 1984), into what is now a very well specified dual process theory of both reasoning and memory (e.g., Reyna, 2004).

Finally here, some theorists (e.g., Bouwmeester et al., 2007; Dayton, 1998; Schnall & Gattis, 1998; Wright, 2006a) provide evidence that the relational aspects of transitive reasoning are quite

well developed from 5 years, but not fully developed until well beyond childhood. Some also report data on age of peak transitive performance. For example, there seems to be a peak at roughly 18 years and a decline from middle age onwards (Hong & Chond, 2001; Viskontas et al., 2005). As with the dual process accounts outlined above, this finding for transitive inference is directly mirrored by recent conclusions from reasoning tasks more generally (Evans, 2003; Kokis et al., 2002). Responding to the above issues, one aim of the present paper was to provide research findings that bridge the gap between children and adults, by working with participants who could be considered just too old to be typical of children and just too young to be typical of adults. According to all relevant theories, such a participant group would exhibit transitive performance at near ceiling (Acuna et al., 2002; Demarais & Cohen, 1998; Wright, 2006b).

2, Established Transitive Phenomena and the B&T Task

One can ask whether typical phenomena found using variants on the B&T task are fully representative of active processes of transitive inference, or simply more general and low level phenomena tied only to the B&T task. For a 5-term series, the principal phenomenon should always have been the B?D inference (Holcomb et al., 1997; Titone et al., 2004). After all, it was to prevent “non-logical solution strategies” about this particular comparison-pair that Bryant and Trabasso brought together procedures from many previous studies into a single task/paradigm (Bryant & Trabasso, 1971; Trabasso, 1977). However, although initially the B?D inference may have been the main phenomenon of interest, other interesting effects were soon noted (for reviews see Allen, 2006; Hummel & Holyoak, 2001). For example, since the year 2000 alone, over 100 research papers have used the finding of a symbolic distance effect as evidence of deductive transitive inference (Bond et al., 2003; Breslow, 1981; Siemann & Delius, 1996; Titone et al., 2004; Trabasso, 1977). This effect can be explained using the convention of 3-step comparison for the A?E case, because in the implied series, there are three items B, C and D between the items being inferentially compared. Likewise, the A?D inference would be a 2-step comparison, the B?D inference a 1-step comparison, and the B:C premise would be a 0-step comparison. The

symbolic distance effect is then defined as the robust tendency for items having greater inferential step to be easier to solve than those with fewer inferential steps (Acuna et al., 2002; Favrel & Barrouillet, 2000; Moses et al., 2006; Wu & Levy, 2001).

The original interest in the symbolic distance effect stemmed from it readily separating an information processing approach to transitive inference from the Piagetian approach. Specifically, Trabasso argued that a Piagetian account of how the premises are coordinated in reaching an inference, predicts a reversed symbolic distance effect, with comparisons of greater inferential step being harder and slower to solve. Because the empirical evidence did not support a reversed symbolic distance effect, B&T advocates concluded that the Piagetian account of transitive inference must be incorrect (but see Breslow, 1981 for a neo-Piagetian account accommodating the standard symbolic distance effect, and Wright, 2006b and Favrel & Barrouillet, 2000 for reversed symbolic distance effects within B&T task variants). Today, the symbolic distance effect is held in such regard that many theorists take it as a stronger indicator of deduction than the B?D inference itself (Acuna et al., 2002; McGonigle & Chalmers, 1992; Wynne, 1998).

The symbolic distance effect is explained in terms of the reasoner representing the entire series in memory and using a mental search strategy in order to respond to pair-wise comparisons. Items further apart are towards the start of an ends-inward search strategy (Breslow, 1981), less confusable (Hummel & Holyoak, 2001), or associated with greater differentials in reinforcement value (Wynne, 1998). Any of these explanations would account for the profile of 3-step comparisons being easiest... 0-step comparisons being hardest. But it is curious that every article finding the symbolic distance effect also seems to have relied on a variant of the B&T task. This raises the possibility that the symbolic distance effect is an artefact of B&T-specific procedures such as the way the training regime forces the reasoner to hold the premise information in memory rather than a natural corollary of transitive inference (Allen, 2006).

A similar argument may be made for other key phenomena. Wynne (1998) argued for the existence of a lexical marking effect. Here, the most valued label is said to be unmarked, because

it is the default adjective for the relational comparative dimension. For example, if the dimension were bigger v smaller then any comparison-pair associated with an item given the unmarked label big more often during training would tend to be resolved more easily than one associated with the marked label small during training (Goodwin & Johnson-Laird, 2005; Hummel & Holyoak, 2001). Thus, A:B and B:C should be easier than C:D and D:E, a finding both predicted from a statistical model and actually found for 5 to 13 year-olds (Bouwmeester et al., 2007). In a similar way, Breslow (1981) argued that the implied series is internalised from both ends towards the centre. Thus, the end two premises are acquired before the middle two premises. However, Wright (1998b) cast some doubt on this assertion as an explanation of how transitive inferences are made, both with child and adult participants. Indeed, in a robust replication with adults, although the expected profile was seen after extensive training, it was actually the inner premises that were learned first (Perner & Aebi, 1983; Wright, 2006b).

A further transitive phenomenon is the serial position effect. This refers to the finding that average performance involving a particular item improves the closer that item is to either end of the series (Bryson & Leong, 2007; Holcomb et al., 1997; Wynne, 1998). In the limit, comparisons involving item-C are generally the hardest, with a U-shaped curve seen for response-accuracy and an inverted-U for RT. This phenomenon may be considered a generalised instance of another effect called the end-anchor effect. The end-anchor effect refers to the common finding that comparisons involving one or both end-items are easier to resolve than those involving neither end-item. Various predictions follow. First, the A?E comparison is easiest to resolve of all (Maydak et al., 1995; Martin & Alsop, 2004; Wu & Levy, 2001). Next, a response to the A:B premise should be easier than to the D:E premise; a response to the A?C inference easier than the C?E inference; and a response to the A?D comparison easier than the B?E comparison (De Lillo, Floreano & Antinucci, 2001; Maydak et al., 1995). Overall, average performance for comparisons involving item-A should be superior to those involving item-E. On this issue, Wright (2006b) used a B&T variant which avoided all short-cut cues and kept training to a level that produced

around 90% overall performance. Wright found that there was no response-accuracy advantage to the A item, even at the end of training. In fact, it appeared that if anything, it is comparisons regarding the E item that lead to more accurate performance. This was interpreted as support for the thesis that the training regime, rather than the fact that transitive inferences were being made, lays behind the end-item effect. Thus, a study not relying on training is crucial to understanding this and the other effects.

3, Response-Times for All Responses v Response-Times for Only Correct Responses

Transitive research uses response-accuracy as an index of performance. Many recent tasks also use RT instead of or additional to response-accuracy (e.g., De Lillo et al., 2001; Goel et al., 2004; Holcomb et al., 1997; Hummel & Holyoak, 2001; cf. Trabasso, 1977; Viskontas et al., 2005). However, just as with claims about who possesses a transitive inference competence and which phenomena demonstrate transitive reasoning, it would appear that all the reported profiles for RT have been found in variants of the B&T task. There is thus a real danger that the B&T task has become synonymous with the phenomenon of "transitive inference" itself.

Most transitive experiments using response-times, take RT to be the times across all responses for a given comparison-pair (Acuna et al., 2002; McGonigle & Chalmers, 1992). Because this use of RT is non-discriminative, we refer to it here as nRT and distinguish it from RT for correct responses only (cRT). However, if we are especially interested in cognitive mechanisms underlying correct transitive inference, then we should focus on cRT rather than nRT (Holcomb et al., 1997). This is because only correct responses tell us about the relative speeds of processes that generate veridical transitive inferences. Indeed, in other cognitive domains we routinely focus only on RTs for correct responses, or circumvent the whole issue by ensuring that incorrect responses are so few that the two alternative strategies essentially become the same (e.g., in the Stroop task –Wright & Wanley, 2003). The problem here is that transitive responses can be both near ceiling and near random chance for different comparisons within the same experiment. For this reason we decided to employ both nRT and cRT analyses. This also allowed us to compare

them to determine whether restricting analyses to cRT only, results in the same profiles across the transitive phenomena as found when nRT is used.

To address all the above issues, we followed suggestions from Wright (2001) and De Lillo et al. (2001) and devised a task based around a 5-term transitive series but that did not necessitate any training to remember premises (Markovits, Dumas & Malfait, 1995). The task used the relational comparison of size (e.g., "bigger than"). Following criteria recently set out by Wright (1998b), all the premise pairs were simultaneously visible on a computer screen, but the various premises were separated spatially. However, in line with some recent tasks using 3-term designs (e.g., Wright, 2006a; Wright & Dowker, 2002), participants only answered one question on any given transitive series, and then the premises were removed and replaced with new premises implying a different series.

Method

Participants

Participants were 58 young adults (36 female) from a college of further education in the West Midlands region of the UK. Their mean age was 17.89 years old. Three participants' data were excluded from analyses presented later. Two did not follow the task instructions and one gave RTs more than 4 standard deviations above the mean for the whole sample.

Materials

The main materials were a program to present premise pairs and collect responses. This was run on a PC compatible portable computer with a 2.4GHz PentiumM processor and a 50cm high-resolution colour display. An external keypad device was used to collect responses. A paper version of the task was also used. Subject to each participant's consent, the procedure was recorded using a tape recorder and microphone. This provided a means of noting any issues raised by the participant or notes made by the experimenter.

Design

The experiment concerned the transitive relation of size, with all questions given in terms of the

unmarked label “bigger than” only. Each item in the transitive series was indexed by colour, as done in most transitive experiments with humans. The experiment itself was based on a repeated measures factorial design, and took response-accuracy, nRT and cRT as the main dependent variables of interest. Stimuli implying a 5-term transitive series were presented via four linearly interlinked premise pairs. This permitted each of the effects to be analysed via combinations of the 10 resultant pair-wise comparisons. For example, the symbolic distance effect could be analysed by grouping the four premises and six inferential comparisons according to inferential step.

The criteria outlined below were each randomised within limits. Each premise pair comprised two identical shapes side by side, differing in size and uniquely indexed by colour. The sizes and size-differentials were determined by earlier piloting, and were those which adult participants were shown to differentiate at between 98% and 100% within 2 seconds of presentation (Acuna et al., 2002). One premise was shown in each quadrant of the screen, with the real distance between the closest part of any two premises not less than 50mm and not more than 110mm in any direction. The distance between the closest part of the items comprising any given premise was between 5mm and 10mm. The height difference between the larger and smaller item of a pair was always 2mm, with their respective heights within a given pair being 31mm and 29mm. Additionally, the largest item in the series had a fixed width of 30mm, with the difference in width for items in a pair a constant 1mm, leading to an absolute and linear gradation of width for the series and a difference in width of 5mm between largest and smallest items in the series. With a minimum viewing distance of 500mm, the result was that the average premise subtended a vertical visual angle of 3.4 degrees and a horizontal of 7.1 degrees. Although the height variable represented categorical relationships, the width variable captured a linear gradation from the A item to the E item. Thus, in response to concerns raised about relative differences versus absolute overall differences (Markovits & Dumas, 1999; Wright & Dowker, 2002), we achieved presentation of stimuli that did in fact reflect a gradation but which participants would likely

interpret as categorical; leading to solution strategies more likely to be deductive.

Procedure

If signed consent was given, the participant was introduced to the testing environment. The experiment was conducted in a quiet room within the college. The computer was set up with a separate keyboard so that the required viewing distance could be achieved. The purposes of the recording equipment was explained to the participant and then turned on only if additional consent was given for this. Every participant gave this consent. Before beginning the task proper, the participant was given a representation of the task on paper. Examples of the screen layout and stimuli were presented in colour, to familiarise each participant with the procedure ahead of the computerised task. The participant was asked at least one question regarding items relevant to the possible 0- to 3-step inferences. A new configuration was then presented with only one question asked, to familiarise the participant with the procedure to follow. The participant was given feedback on his/her answers and any questions or issues raised were discussed. It was also stressed that in the computer task, the participant should try to get each answer correct but also try to respond as quickly as s/he felt appropriate. The latter stage was repeated for four separate paper configurations. Each configuration used here concerned a different shape – rectangles, circles, sticks and triangles. Their purpose is explained below.

On-screen instructions explained the task and the requirements. The experiment was divided into three blocks, with a break between each. Lengths of breaks were determined by the participant, who pressed a key when ready to continue. The composition of the blocks was similar, with each having 40 pairs of displays and asking one question per display-pair. For each display-pair, a screen was first shown, having a question at a designated location at the bottom (following Wright, 1998b, 2006b). The question was of the form “Would A be bigger than B?”. This remained on screen until the participant was ready for the actual stimuli. When the participant pressed a designated key on a separate keypad to signal readiness, the question remained whilst the main part of the screen now introduced the premise information (the

configuration). Here, four item pairs were shown simultaneously on screen. The participant might choose whether to remember the question whilst viewing the premises, or check the bottom of the screen as and when s/he felt necessary. The participant pressed one key to signal "yes" and another for "no". The computer recorded accuracy and RT for the first key-press and then waited for the participant to press the designated key to move on to the next question/configuration.

In order to minimise priming or proactive inhibition/interference effects from one configuration (i.e., one implied series) to another, each block comprised 10 presentations using a given shape, then 10 with an alternative shape and so on for the four shapes. For any given shape, the 10 presentations together assessed performance on each of the four premises and six inferences, doing so once each. Neither the same series nor the same item/colours were used twice in consecutive display-pairs. For 50% of questions about two particular items of a premise/inference, the order of the item referents in the question was reversed ("A>B" v "B>A"). These were distributed across blocks and also distributed across shapes as evenly as possible. This meant that altogether there were 12 questions for each comparison-pair, half of which used each question referent order. The resultant presentations and question referents were randomised ahead of testing, with two such complete randomised orders of the 120 stimuli used to give two different versions of the task. Each participant sat only one of the two versions, determined by chance.

The above design and procedure resulted in a task that used immediate presentation, did not require memorising of premises and asked only one question per series. It could therefore be described as a "non-training single-response task". Altogether the task took between 20 and 30 minutes to complete, including briefing and debriefing.

Results and Discussion

For response-accuracy data, correct responses for each comparison-pair in turn was summed across blocks, question-directions, series and shapes (max = 12). The resultant data are presented as percentages to aid more ready comparison against previous studies. Concerning response-times, the median nRT for each comparison-pair was calculated in a similar way to response-

accuracy. Relying on the median instead of the mean minimised outlier effects on a participant by participant basis. As well as improving homogeneity, this also meant that we did not have to apply any clean-up strategies to our data (contrast McGonigle & Chalmers, 1992). Next, we obtained the cRT dataset by repeating our procedure for only those individual responses that had been correct. A series of Repeated Measures Analysis of Variance (ANOVA) were used to assess the statistical significance of the relevant trends shown in the response-accuracy, nRT and cRT data. Planned contrasts were conducted as appropriate. All tests were two-tailed with an alpha level of 0.05. In the following sections, we present and briefly discuss response-accuracy and RT findings for (a) overall solution of the implied-series; (b) the symbolic distance effect; (c) the lexical marking effect; (d) the inward-acquisition effect; (e) end-anchor effects, (f) the serial position effect; and (g) the critical B?D inference and other 1-step inferences.

(a) Overall Solution of the Implied-Series

Percentages of correct solutions to each of the 10 possible pair-wise comparisons are given in Table 1, with the corresponding cRTs and nRTs given in Table 2. Many theorists (e.g., Bouwmeester, 2007; Bryant, 1998; Halford & Andrews, 2004) intimate that, for humans, if one can solve one B&T task variant one can also solve another to the same extent. Indeed, without subscribing to this view, one cannot maintain the B&T task ever did target the same transitive competence as the classical 3-term task it dominates (Wright, 1998a). But from Table 1 we see that, across all 10 comparison-pairs, the overall response-accuracy was 63.8%. Also, response-accuracy differed from one pair to the next. This was confirmed as statistically significant with a one-way repeated measures ANOVA (F (9, 486) = 57.075, $p < 0.005$, Obs.Power = 1.000).

Thus, performance here was more challenging than for any prominent study using the B&T task; including arguably the most demanding B&T variant to date (Wright, 2006b). This might be surprising, given that B&T variants require memorisation of premises, whereas here all premises were continually available for inspection. It adds support to Wright's (1998a, 2001) contention that, in the generic B&T task, the method of committing the premise information to memory, such

as by over-training participants and using interlinked ascending/descending premise training, might "induce" rather than simply "allow" transitive solutions (Martin & Alsop, 2004; Russell et al., 1996; Stromer et al., 1993). Of course, the lower overall performance reported here might simply be a cohort effect. However, against this argument, we note that all our participants were in advanced level education (A' Levels) and other researchers have found that participants at this age are approaching their peak for transitive reasoning (Demarais & Cohen, 1998; Hong & Chond, 2001; Viskontas et al., 2005).

Insert Table 1 about here.

Turning to RT, neither the B&T task nor Wright's variant on the classical task make explicit claims about nRT versus cRT. However, a generic cognitive view would hold that some correct answers are due to the participant considering the information just a little more carefully, or due to processing times for mentally sifting and selecting the logically-necessary information, hence taking longer. Conversely, some wrong answers would stem from a loss of concentration, premature responding or timeout; hence faster responses. Although neither set of strategies would be the sole basis of either correct or wrong responses, any tendency towards these respective strategies would lead to nRT being slightly speeded relative to cRT. Data for nRT are summarised in Table 2 (Top) with cRT summarised in Table 2 (Bottom). We first note that nRT was an average of 299ms faster than cRT. A two-way ANOVA was conducted, with factors of RT-Type and Comparison-pair. This showed a statistically significant difference between the two RT measures (F (1, 54) = 19.247, $p < 0.005$, Obs.Power = 0.991). As with response-accuracy earlier, the RT for the 10 pair-wise comparisons differed from one another; with the difference statistically significant both for nRT and cRT (F (9, 486) = 11.701, $p < 0.005$, Obs.Power = 1.000). Table 2 suggests a rather complex interaction effect, with middle premise pairs and the 1-step inferential pairs showing bigger differences between nRT and cRT. This interaction was statistically confirmed (F (9, 486) = 4.559, $p < 0.005$, Obs.Power = 0.999).

Perhaps with the exception of Holcomb et al. (1997), transitive inference researchers tend to

intimate that it is sufficient to use nRT as the index of speed of responding. However, our findings here show that, not only does nRT lead to over-favourable estimates of speed of decisions during transitive reasoning, but the amount of speed inflation depends on which particular comparison-pairs are involved. This conclusion is borne in mind for the various transitive phenomena in the following sub-sections.

Insert Table 2 about here

(b) The Symbolic Distance Effect

The original B&T claim (Trabasso, 1977) was that retrieval from the kind of mental array set up by participants faced with the B&T task, should result in worst performance for those items most close together in the mental representation of the implied series, because their closeness makes them most difficult to discriminate between. Accuracy then improves the further away the items to be inferentially compared (see also Wynne, 1998; Wu and Levy, 2001). Because this profile is almost invariably accepted both in human and comparative research, we refer to it as the "Standard Symbolic Distance Effect". The last column of Table 1 summarised our response-accuracy data by inferential step. The general trend seems in line with Trabasso's (1977) original assertion. A one-way ANOVA statistically confirmed the overall tendency towards a symbolic distance effect (F (3, 162) = 49.026, $p < 0.005$, Obs.Power = 1.000). But crucially, Table 1 shows that the 0-step pairs were not in line with this overall effect. Additional analyses by way of difference-contrast showed that 3-step inferences were superior to all the other inferential steps combined ($p < 0.005$). The 2-step inferences were then superior to 1-step inferences. But the tendency for 0-step inferences actually to be superior to 1-step and 2-step inferences combined was statistically significant (each $p < 0.005$).

For RT, the B&T view again holds that we should find a standard symbolic distance effect, with outermost pairs attracting the fastest times (Acuna et al., 2002). The last column of Table 2 summarised for the symbolic distance effect. We first note the similar difference for nRT versus cRT to the difference between them that we noted for all 10 pairs (309ms). Of importance here,

the pattern of RTs did seem to reflect some variations with inferential-step, but these appeared stronger for cRT than for nRT. We assessed these trends using a two-way ANOVA, with factors of RT-Type and symbolic-distance. The main effect of RT-Type was statistically significant (F (1, 54) = 16.114, $p < 0.005$, Obs.Power = 0.976). Symbolic-distance was also statistically significant (F (3, 162) = 11.710, $p < 0.005$, Obs.Power = 1.000).

To assess the symbolic distance effect in more detail, difference-contrasts were computed separately for nRT and cRT. For nRT the contrast between 0-step pairs and 3-step pairs did not approach statistical significance and nor did the difference between 2-step and the previous steps combined (each $p > 0.100$). However, the difference between 1-step and the three other inferential-step categories was statistically significant, both taken individually and combined (each $p < 0.005$). Thus, any contribution of nRT to the significant overall main effect of symbolic distance in the two-way ANOVA was driven mainly by the much longer nRT for the 1-step inferences, with no systematic symbolic distance effect from other inferential steps.

It is tempting to conclude the symbolic distance effect is abolished altogether by avoiding the B&T task. However, before reaching this conclusion, we need to determine whether its absence for nRT is mirrored by cRT. Table 2 and difference-contrasts analogous to those for nRT, together confirm that for cRT, 3-step inferences were significantly faster than 2-step inferences, and 2-step inferences faster than 1-step inferences (each $p < 0.010$). As found regarding response-accuracy, the 0-step responses were significantly faster than the other inferences on an individual basis (each $p < 0.010$). There was only one exception to this pattern, which was the contrast between the 0-step and 3-step responses ($p > 0.100$). Thus, unlike nRT, cRT shows a symbolic distance effect, although not the standard effect usually claimed from B&T tasks.

Comparing the top portion against the bottom portion of Table 2, shows a tendency for the difference between the two RT-Types to increase when two or more premises must be coordinated in order to deduce the response (i.e., 1-step and 2-step inferences), and to decrease greatly when no inference was logically required (i.e., for 0-step and 3-step responses). The suggested

interaction was confirmed in our above two-way ANOVA analysis by a significant interaction effect between RT-Type and symbolic-distance (F (3, 162) = 3.552, $p < 0.050$, Obs.Power = 0.779).

So, whilst there is a symbolic distance effect for response-accuracy, it is not the standard symbolic distance effect the B&T task predicts. Rather, the effect holds only for inferential comparisons, with the 0-step taught pairs responded to more accurately than any other pairs apart from the 3-step end-pairs. Then for RT, when we use the index of nRT used in most transitive research but within a task that rules out passive seriative generalisation, the symbolic distance effect is abolished, with only the 1-step inferences (i.e., those most likely to require deduction here) showing longer nRT. But if we focus only on correctly given responses, then our cRT index now mirrors the non-standard symbolic distance effect we found for response-accuracy (i.e., it excludes the taught pairs). In line with our present findings, Wright (2006b) confirmed an overall symbolic distance effect for accuracy and nRT, on a B&T task variant; but noted that the 0-step versus 1-step comparisons did not fit the standard symbolic distance effect (accuracy - 88.5% v 88.3%; nRT - 879ms v 921ms). An actual 0-step superiority was reported for children both on 3-term and 5-term transitive tasks (Wright, 1998b, 2006a; Wright & Dowker, 2002). This profile can even be seen within Bryant and Trabasso's (1971) seminal paper and in Wu and Levy's (2001) computational data.

(c) The Lexical Marking Effect

The lexical marking effect was assessed as the difference between the two premises at the large (unmarked) end of the implied series versus those at the small (marked) end. These were A:B and B:C versus C:D and D:E respectively. Table 3 summarises this effect along with other effects and evaluation of whether each effect fits with the B&T task. For lexical marking, Table 3 suggests that for response-accuracy the effect was actually in the reverse direction to predictions from the B&T task (Wynne, 1998). A one-way ANOVA statistically confirmed this tendency (F (1, 54) = 28.750, $p < 0.005$, Obs.Power = 1.000).

For RT, Table 3 shows that nRT was some 162ms faster than cRT overall. Roughly the same difference occurred both at the unmarked end and the marked end of the series. Table 3 also suggests a slight tendency towards a lexical marking effect for nRT and cRT (85ms and 75ms, respectively). We analysed the statistical significance of these data in the same way as for response-accuracy earlier, apart from adding a second factor in the ANOVA with two levels corresponding to RT and cRT. In the resultant two-way ANOVA, the main effect of RT-Type was statistically significant (F (1, 54) = 9.164, $p < 0.005$, Obs.Power = 0.844). However, the main effect of lexical category failed to reach statistical significance (F (1, 54) = 2.142, $p > 0.100$, Obs.Power = 0.301). Any two-way interaction also failed to reach statistical significance (F (1, 54) < 1.000 N.S.).

So, we did find a lexical marking effect for response-accuracy but this was in the reverse direction to expectations from recent theorising, the B&T task and at least one recent experiment (see respectively, Bouwmeester et al., 2007; Goodwin & Johnson-Laird, 2005; Hummel & Holyoak, 2001). Also, regarding RT, there was no reliable lexical marking effect for either nRT or cRT.

Insert Table 3 about here

(d) The Inward-Acquisition Effect

Wynne (1998) presented alternative mathematical models consistent with the B&T assertion that the implied-series is built up from the ends-inwards. Using a variant of the B&T task with adults, Wright (2006b) replicated the finding that outer premises are responded to more accurately and faster than inner premises by completion of training. However, in the same experiment, it was also found that, early in training, inner premises are actually acquired better than the two outer premises (see also Perner & Mansbridge, 1983). We therefore sought to clarify using a task that encourages deductive solutions without inadvertently inducing seriative generalisation. The summary in Table 3 suggests better performance for the outer pairs compared to inner pairs. A one-way ANOVA showed this tendency to be statistically significant (F (1, 54) = 14.956, $p <$

0.005, Obs.Power = 0.967). The finding that inner premises have lower response-accuracy is in agreement with Acuna et al. (2002) and Breslow (1981). It implies the initial profile in Wright (2006b) stemmed from the influence of a different mode of reasoning to the final profile in that experiment; which Wright argued were associative and deductive, respectively.

For RT, Table 3 shows outer premises were responded to around 269ms faster than inner premises overall. Also, precisely the same difference in inward-acquisition was evident for nRT compared to cRT; although as usual, nRT tended to be the faster index. These trends were assessed using a two-way ANOVA analogous to that for lexical marking. The difference between nRT and cRT was statistically significant (F (1, 54) = 9.164, $p < 0.005$, Obs.Power = 0.844). The main effect of inward-acquisition was also statistically significant (F (1, 54) = 32.468, $p < 0.005$, Obs.Power = 1.000). There was no statistically significant two-way interaction (F (1, 54) < 1.000 N.S.), showing that the contrast between inner and outer premises was identical for nRT and cRT.

Both the response-accuracy and RT profile for inward versus outward premise superiority are in line with assertions from the B&T task (Breslow, 1981). Note, the main difference between inner and outer premises is that outer premises involve one of the end-items. It therefore seems that the inward-acquisition effect may be in line with B&T task predictions and yet still be an artefact of the B&T task. For example, the reasoner may simply tend to check whether one of the items is in an end premise, and if not, then proceed to consider one of the inner premises instead. Indeed, this explanation can also account for the non-standard symbolic distance effect found earlier for cRT: Any inferential comparison involving an end-item can be solved either deductively by considering and coordinating the two antecedent premises, or non-deductively by recognising that item-A will always be larger of any comparison and item-E will always be smaller in any comparison.

(e) End-Anchor Effects

B&T advocates generally maintain that the first and last items within the implied series act as end-anchors, with the series built up from these items (Breslow, 1981). Our finding that outer premises are responded to faster and more accurately than inner premises lends some support to

this assertion. Item-A is usually claimed to be the uniquely unmarked item in that it is never given the marked label in premise or inferential comparisons. However, in practice it is not unambiguously unique because each of items-B, -C and -D are also given the unmarked label in at least some of their comparisons (Goodwin & Johnson-Laird, 2005; Wright, 2001). The fact that all items except item-E are given the unmarked label in at least some comparisons, may be taken together with the typical assertion that the reasoner thinks primarily in terms of unmarked relations (this is why they are termed unmarked in the first place – Wynne, 1998). This leads to the posit that item-E is in some practical sense even more unique than item-A, because it is the only item never a candidate for a decision (i.e., the reasoner never has to indicate that it is larger of any comparison).

Figure 1 (Top) depicts response-accuracy on the three comparison-pairs involving the A end-item versus those for the E end-item, but excluding the A?E pair which was common to both end-items. It can be seen that for each pair of bars in the figure, the E comparison was indeed superior to the A comparison. This trend was assessed via a two-way ANOVA. The factors were end-anchor (A v E) and comparison-pair. The main effect of end-anchor was statistically significant ($F(1, 54) = 78.414$, $p < 0.005$, Obs.Power = 1.000). The main effect of anchored-comparison-pair was also statistically significant ($F(2,108) = 17.216$, $p < 0.005$, Obs.Power = 1.000). Figure 1 shows a tendency for the 0-step comparison-pair to have an advantage over the 1-step and 2-step pairs that involve item-E, but the reverse pattern regarding item-A. This was confirmed by a statistically significant two-way interaction ($F(2, 108) = 5.742$, $p < 0.005$, Obs.Power = 0.858).

Insert Figure 1 about Here

The end-anchor effect for RT is also depicted in Figure 1. This effect was examined separately for nRT and cRT. Taking nRT first, Figure 1 (Middle) shows little overall difference between responses involving item-A and those involving item-E. There was, however, a difference according to comparison-pair, with the 1-step pair attracting the slowest RT. A two-way ANOVA analogous to that for response-accuracy showed no overall main effect of end-anchor ($F(1, 54) <$

1.000 N.S.). Although not in line with the B&T task, this finding mirror the profile found by Wright (2006b), on a B&T-variant intended to remove as many non-deductive routes to solution as possible. Our analysis revealed the difference between the pairs to be statistically significant (F (2, 108) = 13.261, p < 0.005, Obs.Power = 0.997). Figure 1 shows a tendency for item-A to have an nRT advantage for the close pairs, turning into an item-E advantage for the far pairs. This interaction was statistically significant (F (2, 108) = 3.425, p < 0.050, Obs.Power = 0.632).

For cRT, Figure 1 (Bottom) shows a larger overall difference between comparisons involving the A item and those involving the E item, with E tending to result in faster performance. A two-way ANOVA showed that unlike for nRT, the main effect of end-anchor was statistically significant (F (1, 54) = 15.481, p < 0.005, Obs.Power = 0.972). The main effect of pair was also statistically significant (F (2, 108) = 17.220, p < 0.005, Obs.Power = 1.000), as was the two-way interaction (F (2, 108) = 3.449, p < 0.050, Obs.Power = 0.636).

Regarding RT, when we use the nRT format favoured by most B&T studies, we find no robust end-anchor effect. However, when we rely on cRT instead of nRT, we do find an end-anchor effect, but contrary both to earlier studies based on nRT only (e.g., Acuna et al., 2002) and also cRT studies (e.g., Holcomb et al., 1997), the cRT end-anchor advantage is for item-E (i.e., the marked end-anchor) rather than item-A (i.e., the unmarked end-anchor). Our cRT findings were closely matched by our findings for response-accuracy, with the marked end-anchor again having an advantage over the unmarked end-anchor (contrast Bouwmeester, 2007; Stromer et al., 1993; Trabasso, 1977). Our claim of an item-E advantage is actually implicitly supported by a number of existing B&T studies, even though the papers concerned rarely acknowledge and discuss this aspect of their data (e.g., see Russell et al., 1996). Wright (2006b) has reported that the tendency for item-E to be the pivot of the entire series is present even in studies based around fully randomised B&T tasks, if we measure performance early in training.

To focus briefly on end-anchor premise pairs, the B&T task predicts that the A:B premise would be responded to fastest and most accurately (Breslow, 1981; Trabasso, 1977). However, as can be

seen from Figure 1, the D:E premise actually holds an advantage over A:B both for response-accuracy and cRT; although it is reversed for nRT. But this D:E advantage is denied in most theories of transitive reasoning and hence not model led in computer simulations (e.g., see De Lillo et al., 2001; Wu & Levy, 2001). Yet a D:E advantage can be readily observed in data summaries of many studies using the B&T task: It emerged in Moses et al. (2006) with young adults, in all three experiments of Siemann and Delius (1996) with adults, and all three tasks of Russell et al. (1996) with children.

(f) The Serial Position Effect

According to the serial position effect, performance should get slower and less accurate as we move away from either end of the series toward the middle; and performance should favour comparisons involving the unmarked item-A over the marked item-E (Bryson & Leong, 2007). This effect was indexed by averaging all four premise/inferential pairs that involved item-A. The same was then done for the four pairs involving item-B, including pair B:A, and so on for items C, D and E. The response-accuracy profile across the five items is shown in Figure 2 (Top). This shows that item-A attracted lowest rather than highest accuracy, with B, C and D higher than A but similar to each other (although systematically decreasing), and item-E attracting the highest performance. The statistical significance of these differences was assessed using a one-way ANOVA, with five levels. This yielded a significant effect of item position ($F(4, 216) = 66.879$, $p < 0.005$, Obs.Power = 1.000). Contrast analyses indicated that item-C yielded significantly lower accuracy than B, with item-D lower than both (each $p < 0.005$). Less surprisingly given our earlier findings, item-A was lower than B, C and D; with item-E higher than them all (each $p < 0.005$).

The serial position effect for RT was calculated in the same way as response-accuracy. Figure 2 (Bottom) summarises the nRT and cRT profile from item-A to item-E. For nRT, this shows slowing from item-A through to C, and then speeding up to item-E. Also, E was responded to faster than A overall. For cRT, the slowest performance was once again in the middle of the series but this time, item-D had a slightly slower cRT than C. As for nRT, the cRT for item-E was

fastest, but this time item-A was responded to more slowly than item-B as well. A two-way ANOVA showed that RT-Type was statistically significant (F (1, 54) = 19.247, $p < 0.005$, Obs.Power = 0.991). Item-position was also statistically significant (F (4, 216) = 6.854, $p < 0.005$, Obs.Power = 0.993). The difference between nRT and cRT did not seem to follow any simple pattern. This difference was largest for item-A and smallest for item-E but there was no obvious pattern for B, C and D. Nevertheless, this A versus E difference facilitated a statistically significant two-way interaction (F (4, 216) = 5.494, $p < 0.005$, Obs.Power = 0.975).

So, we did not find the U-shaped serial position effect B&T theorists predict for response-accuracy. But the finding that item-A had a disadvantage rather than an advantage compared to item-E, with item-E superior to the middle three items, is in line with recent findings by Wright (2006b), although it contrasts with other previous findings (e.g., Hummel & Holyoak, 2001; Wynne, 1998). Hummel and Holyoak's (2001) computational model of transitive reasoning does not explicitly cover this serial position effect. Yet, the method they employed for placing the premises within an integrated spatial array actually results in each item from each premise being given an absolute location; this reducing to the standard serial position effect. In partial agreement, we concede a "restricted serial position effect", as long as we ignore the two end-items. We found stronger evidence of the inverted-U-shaped serial position effect for nRT, with slowing from each end towards item-C. However, this effect was ambiguous because the fastest performance was not at item-A but rather at item-E. A similar profile was found for cRT, apart from slowest performance now being at item-D. In both cases, the RT effect was driven more by item-E than by item-A.

(g) The Critical B?D Inference and Other 1-Step Inferences

The final set of analyses concerned each of the 1-step inferences and its corresponding antecedents. These were for the critical B?D inference, A?C inference and C?E inference, respectively. Beginning with B?D, we see from Table 3 that for this inference, B?D accuracy was lower than the mean of B:C and C:D. A one-way ANOVA was conducted concerning B?D. This

had three levels, the first two of which corresponded to the two antecedents B:C and C:D, and the last corresponding to the B?D inference. This yielded a statistically significant main effect of comparison-pair (F (2, 108) = 9.448, $p < 0.005$, Obs.Power = 0.977). Difference-contrasts showed that B?D was significantly lower than its two antecedents combined ($p < 0.005$).

For RT, Table 3 shows that nRT for B?D was slower than the mean of its antecedents. A similar, if more marked, profile is shown for cRT. A two-way ANOVA tested the reliability of these trends. The first factor being RT-Type and the second factor being comparison-pair. The difference between nRT and cRT was statistically significant (F (1, 54) = 17.874, $p < 0.005$, Obs.Power = 0.986). The differences between the pairs was also statistically significant (F (2, 108) = 13.562, $p < 0.005$, Obs.Power = 0.998). There was a statistically significant two-way interaction effect (F (2, 108) = 4.181, $p < 0.050$, Obs.Power = 0.725). Difference-contrasts for nRT indicated no statistically significant difference between B:C and C:D ($p > 0.100$) but a significant difference between the B?D inference and its antecedents combined ($p < 0.010$). Corresponding contrasts for cRT indicated that premise C:D had a longer cRT than B:C, and the B?D inference had a longer cRT than its antecedents, whether they were taken individually or averaged (each $p < 0.050$).

For the first of the non-critical 1-step inferences, Table 3 shows that the A?C inference was again less accurate than its antecedents. This was confirmed using a one-way ANOVA, by a statistically significant difference between the comparison-pairs (F (2, 108) = 40.043, $p < 0.005$, Obs.Power = 1.000). Difference contrasts then confirmed that A?C was less accurate than its two antecedents combined ($p < 0.005$). Table 2 also shows that A?C was slower than either of its antecedents, with premise A:B responded to much faster than B:C. This pattern held regardless of whether we relied on nRT or cRT, but as before, cRT resulted in longer times. A two-way ANOVA showed the main effect of RT-Type to be statistically significant (F (1, 54) = 19.759, $p < 0.005$, Obs.Power = 0.992), as was the main effect of comparison-pair (F (2, 108) = 23.126, $p < 0.005$, Obs.Power = 1.000). The two-way interaction was statistically significant (F (2, 108) =

8.314, p < 0.005, Obs.Power = 0.959). Difference-contrasts for nRT showed that A?C took significantly longer than its antecedents combined (p < 0.005). For cRT, the A?C inference was again longer than its antecedents combined (p < 0.005).

Considering response-accuracy for the C?E inference, Table 3 shows that, unlike the other two 1-step cases, C?E fell between its two antecedents but was very much closer to D:E. The result was that for this comparison only, the inference was actually superior to the mean of its antecedents; which is in line with item-E being the pivot of the implied series plus items further apart being easier to discriminate between. A one-way ANOVA showed the overall difference between the comparison-pairs to be statistically significant (F (2, 108) = 64.597, p < 0.005, Obs.Power = 1.000). Difference contrasts showed that C?E was significantly higher than its mean antecedents (p < 0.005). This is in line with our conclusion above, that item-E is unique within the implied transitive series, and hence any comparison against this item can be correctly given without the need to coordinate the antecedents deductively (Wright, 2001).

Then, for RT, Table 3 fits the more usual pattern of the C?E inference being slower than its mean antecedents. However, from Table 2 we saw that the C:D premise was responded to with a similar cRT to the C?E inference itself. Furthermore, C:D was markedly slower than D:E. This profile is again in line with our claim that responses involving item-E do not require consideration of any interlinked information. A two-way ANOVA showed a statistically significant main effect of RT-Type (F (1, 54) = 6.221, p < 0.050, Obs.Power = 0.688) and of pair (F (2, 108) = 8.139, p < 0.005, Obs.Power = 0.955). The two-way interaction approached but did not reach statistical significance (F (2, 108) = 2.791, p > 0.050, Obs.Power = 0.539). For nRT, difference-contrasts showed that the antecedents did not differ significantly from each other, and the slower C?E inference compared to the mean of its antecedents also was not statistically significant (each p > 0.050). Things were slightly different for cRT. Now the difference contrast showed that the D:E antecedent was responded to significantly faster than C:D as well as faster than the C?E inference (p < 0.005); with this causing the C?E inference to be slower than the mean of its antecedents but

this difference only reaching marginal statistical significance ($p = 0.060$).

The findings for the three 1-step inferences versus their respective mean antecedents are quite clear. For the critical B?D case, the much lower response-accuracy and slower cRT for the inference compared to its antecedents, suggests that the critical B?D inference was being reached via deduction (Wright, 2006a, 2006b; Wright & Dowker, 2002). Quite surprisingly, despite the A?C inference being influenced by the end-anchor effect from item-A, the response-accuracy and cRT profiles were nevertheless similar to that for the critical B?D inference. Then for the C?E case, the greater uniqueness of the item-E end-anchor compared to the item-A end-anchor, led to the profile shown for A?C and B?D cases, holding for RT (particularly cRT) although it did not quite hold for response-accuracy. Thus, for A?C, deduction still played a part but this time alongside some non-deductive strategy, with the C?E case suggesting deduction may have played little part in the C?E inference. Because gauging the relative contributions of the deductive and non-deductive strategies is difficult for the A?C case and the C?E case, we would suggest that only the B?D inference is relied on as any basis for conclusions about whether a particular group possesses or does not possess a deductive competence for transitive inference.

General Discussion

Before discussing the implications of our findings to our main goal, we consider the findings in relation to our three specific aims as we initially set out. First, this experiment avoided the need for the reasoner to learn and retain the premise information by removing the need for repeated training. However, removing these memory demands did not render our 5-term task easier than the B&T task. Rather, both our overall performance and the critical B?D inference were much closer to the 50% chance level than to the perfect performance one might expect of young adults (63.8% and 55.1%, respectively). So, when we ensure the design and procedures avoid as many short-cut cues to the transitive response as is feasible, and we also avoid a training regime likely to induce rather than assess transitive responses, even adults presumably near their peak in cognitive ability do not perform as well as B&T task studies report for 4 year-old children. One

implication is that, in contrast to the currently dominant view in transitive research (e.g., Bryant, 1998; Halford & Andrews, 2004), transitive reasoning here which is likely to be deductive particularly for the critical comparisons, is not the primary mode of reasoning indexed using most previous B&T tasks (see Markovits & Dumas, 1992; Markovits et al., 1995; Wright, 2006a, b).

Second, to help review the many findings for response-accuracy, non-discriminative RT (nRT) and RT for correct responses only (cRT), we included three evaluative columns in Table 3 (one after each respective measure). The evaluative columns showed that very few B&T task response profiles typically taken as indicating logic, deduction or rationality, remained present on our non-B&T task. Most profiles were either absent or completely reversed for response-accuracy and our two RT measures. Our present finding that only a minority of B&T phenomena appear when we use a task much more likely to encourage participants to rely on their deductive competencies, is further evidence that many of the usual B&T phenomena stem more from the procedures used in those tasks than from the deployment of deduction (Markovits et al., 1995; Wright & Dowker, 2002).

Third, our findings indicate that nRT measures lead to different response profiles to those based on cRT; with cRT supporting more B&T profiles than nRT. This is a pertinent finding, especially given that most B&T studies have neglected to use the more valid cRT index of RT. Those B&T phenomena that are found both on the present task and on B&T tasks more generally, were without exception the ones for which the correct solution could easily be reached by considering only one of the two premises logically sufficient for that conclusion. In other words, they are the ones that could be solved via a less demanding and non-deductive strategy. We should therefore be cautious about using response profiles such as end-anchor effects and the standard symbolic distance effect as direct evidence that a deductive mode for transitive inference is being utilised (contrast Bryant, 1998; Halford & Andrews, 2004; McGonigle & Chalmers, 1992; Trabasso, 1977; Wynne, 1998).

We now turn to our main goal of integrating B&T findings with the present findings. Transitive

research has been heavily influenced by the assumption that the only way to solve for inferential questions is to construct the entire linearly-interlinked series implied by the premises. This assumption is in the child and adult reasoning literature, in the animal literature and even in much of the computational literature (e.g., Acuna et al., 2002; De Lillo et al., 2001; Holcomb et al., 1997; Hummel & Holyoak, 2001; Martin & Alsop, 2004; Siemann & Delius, 1998; Trabasso, 1977). As an illustration, Schnall and Gattis (1998) conclude that effects such as the symbolic distance effect derive from the necessity of constructing some sort of mental array in long-term memory. Similarly, Halford and Andrews (2004) state "The process of constructing the ordered set representation is an important part of the reasoning process, because it is there that the transitivity principle has to be applied." (Halford & Andrews, 2004, pp.126).

By "transitivity principle", Halford and Andrews allude to the target deductive competence. Chapman (1999) concedes that the representation of the implied series in memory (Halford & Andrew's "ordered set representation") may indeed be important. However, Chapman argues that what is even more important are the cognitive mechanisms whose operations result in this mental representation in the first place. For example, it may well be that transitive responses are deduced via the reasoner constructing and reading off from mental models (Bara et al., 2001; Goodwin & Johnson-Laird, 2005; Hummel & Holyoak, 2001). However, what we really need to know is what knowledge, experience or ability leads the reasoner to choose to set up mental models in the first place? Also, what is the basis of the reasoner's realisation that constructing and reading off from such models might solve transitive problems anyway? We would argue that it is these abilities that approximate to any deductive competence in adults. It is crucial to note that none of these abilities are to be gleaned from the mental representation of the entire transitive series itself.

The above point notwithstanding, in our task here, constructing the entire series was in any event never logically necessary. The generally-applicable strategy was constructing temporarily only that part of the implied series necessary to find the answer desired. But although this renders Halford and Andrew's (2004) account unlikely to have been the route to solution for our task here

(Wright, 1998a), the account may yet hold for the B&T task. B&T tasks typically require all 10 possible comparisons to be reported tens of times or even hundreds of times over; with training and test taking anything between around 30 minutes and over 12 months, depending on participant group (McGonigle & Chalmers, 1992; Russell et al., 1996; Wright, 2006b; Wynne, 1998). In that case, it does make sense for reasoners to generate the entire implied series and give their answers by comparing the position of the items from the entire series. But then test performance would say very little about whether the reasoner used a deductive competence, because any deduction would have played its part at a point before testing began (Wright, 1998b, 2006b).

As well as the test procedure, Wright (2001) argued that the method of training to ensure premise retention might induce reasoners to gradually build up the implied series. It was further argued that typical B&T phenomena, such as the symbolic distance effect, reflect the series that has been built up in memory, but are not relevant to the actual use of deduction. On this view, removing the need for training whilst also avoiding repeated questions about a given series, should lead to abolishment of such effects. But for symbolic distance, we abolished it for the usual but less valid B&T nRT measure, only to see it re-emerge for our more valid cRT measure (although it was a non-standard effect that excluded the premise pairs). We might take this as evidence that, even when we avoid extensive premise training, the reasoner nevertheless constructs an integrated mental array and generates transitive inferences from this array, apart from the premises which can be reported perceptually. But in our task here, a symbolic distance effect could arise from a general solution strategy, rather than from mental representation of the entire implied series. Indeed, given that each of the many transitive questions we gave, was based on a different implied series, it is doubtful that reasoners here would have expended cognitive effort extracting the entire transitive series just to answer a single question. However, they could reduce overall load by identifying the end-anchors for each series. Then, if the end-anchor did not permit the inferential comparison to be made, the reasoner would go on to consider whether the

particular premise being focussed on was the one next to an end-anchor that had already been identified. A strategy such as this would lead to a profile resembling the symbolic distance effect. But it would exclude the premises because the solutions to these can be seen without considering any other item. This explanation readily accounts for why the 3-step comparisons were solved as readily as the directly perceivable relationships within the premise pairs: The reasoner started with a view of all the premises, which tended to permit the end-anchors to pop out before the focus was shifted to individual pairs. Further research is required to determine whether our symbolic distance effect (and other effects like inward-acquisition of premises) are due more to entire-series construction or more to the kind of strategy suggested here. For now it is sufficient to note that such effects should no longer be taken as unambiguous indicators of the deployment of deduction. Indeed, only the 1-step inferences are likely to involve deduction, with only one of these (here B?D) critical as an indicator.

Our above account can explain why the few B&T phenomena we found, were evident on our non-B&T task. We now consider why they might be far more evident on B&T tasks themselves. Wright (2001) argues that when extensive training is used, the premises are stored verbatim first, and then the entire series is gradually generalised associatively/passively (Brainerd & Kingma, 1984). Conscious strategies are then used to report “gist” from the generalised series (Brainerd & Reyna, 1992). The point here is that, initially, the premise information is held separately from the generalisations stemming from that information, a posit for which there is some support both from human studies and comparative studies (Eichenbaum, 2001; Fernandez & Tendolkar, 2001: Reyna, 2004; Titone et al., 2004). The verbatim trace supports fast and accurate decisions about premises, and the generalisation trace is more durable and supports the standard symbolic distance effect (Brainerd & Reyna, 1992). This conceptualisation leaves room for the verbatim trace to be coordinated in a one-off manner in order to reach a given transitive solution deductively. In this event, it is reasonable to assume that, should this route be taken repeatedly, it becomes more and more economical to store the result, which would then enhance the seriative generalisation

process (Goodwin & Johnson-Laird, 2005; Reyna, 2004; Wright, 2001). Thus, again we reach the view that without careful theorising, it is problematic to assume that B&T tasks readily distinguish between deductive and non-deductive means of constructing the implied series.

Our findings and the above discussion clearly point to two general modes for reaching transitive solutions. One for situations likely to be one-off, and the other for situations likely to be repeated over and over again: Or one for new or novel situations and the other for heavily memorised information. Or even one for conscious/wilful inferences and the other for unconscious/automatic inferences (for similar conceptions see Bryson & Leong, 2007; Martin & Alsop, 2004; Schnall & Gattis, 1998; Reyna, 2004). We believe our present non-training single-response task calls for the reasoner to engage in deduction, although we concede that for comparisons involving an end-item, a non-deductive strategy might actually take primacy over any deductive competence (Wright & Dowker, 2002).

With the B&T task, the method of ensuring premise retention and the procedure of asking for repeated answers from a single implied series, together tend both to induce the implied series AND to persuade reasoners to engage in a strategy of mentally scanning the internal representation of the series (Stromer et al., 1993). Siemann and Delius (1996) show that B&T tasks can indeed lead to passive generalisation of the transitive series. These researchers presented a transitive experiment as part of a computer game, and found that their most successful adult participants reached up to 100% accuracy without becoming conscious of the fact that the game had presented them with pair-wise alternatives constituting a 6-term transitive series. The summarised data in Titone et al.'s (2004) B&T task, indicate that their adult participants actually improved their performance from testing immediately following completion of training to a further test session with no intervening training. This implies that participants rehearsed and consolidated the series even when no further training was being given. The competence assessed by Siemann and Delius (1996) and Titone et al. (2004), then, is likely to be associative and memory-based. Whether or not one accepts this evaluation, we would argue that it is unsafe to

make statements about deduction from such studies; and yet in the B&T task, such designs appear to be the rule rather than the exception.

Our distinction here between associative and deductive routes to transitive responding is not just tied to transitive inference, but is also in line with a number of recent dual process theories of reasoning more generally (Evans, 2003; Ferreira et al., 2006; Kokis et al., 2002; Reyna, 2004). For instance, in Evans' (2003) dual process theory, the distinction is between "System 2" which is sufficient for deductive inference-making, and "System 1" which is not sufficient for deductive inference-making but may still cause associative links permitting responding that sometimes resembles inferences. Evans (2003) summarises evidence that indicates system 2 (deductive) tends to lead to much slower decisions than system 1 (associative), and is more taxing on linguistic and working memory processes (see also De Neys & Glumicis, in press). In line with this assertion, the average nRT for the present task was some three times slower than found in the B&T variant used by Wright (2006b).

Like Evans' (2003) dual process theory, Ferreira et al. argue that the associative mode is based on generalisations or heuristics, whereas the deductive mode is based on consciously-controlled or rule-based strategies. However, Kokis et al. (2002) and Ferreira et al. (2006) independently note that although dual process theories seem quite compelling, there is as yet fairly little empirical evidence in support of such theories. Our present findings would seem to add to the mounting support for dual process theories. Ferreira et al. (2006) also provide additional evidence by way of four experiments concerning inductive and probabilistic reasoning. They found that priming affected the associative mode but not the deductive mode; whereas secondary tasks, change of reasoning context, or giving unrelated training to induce formal reasoning, each affected the deductive mode but not the associative mode.

Ferreira et al. concluded that the two modes are not competitive, but rather they can be almost totally independent of each other. This contrasts with Evans (2003) who reviewed evidence pointing to an interaction between the associative and deductive modes, whereby the use of the

deductive mode acts to inhibit the utility of the associative mode. But our present results, taken in conjunction with those of Wright (2006b) permit one further postulated relationship. That is, it is possible that the associative and deductive modes derive largely from separate functional systems (Ferreira et al. 's independence); but also that these systems interact dynamically depending on the reasoning task (Evans' suppression). For example, as argued earlier, repeated exposure to the same premises and test questions would lead both the associative and deductive modes to generalise the entire implied transitive series (i.e., support one another). Such a dynamic interplay may serve an adaptive advantage; permitting the more powerful deductive mode to take the lead in decision-making when the context permits (e.g., when there is more time or fewer exposures to premises). However, the deductive system will not be as beneficial when there are constraints on decision-making (e.g., when time is an issue or the transitive solution is sought simultaneously with another train of thought – see Ferreira et al., 2006). Here, we may give increasing primacy to the older associative system, which will tend to generate less demanding and much faster decisions but at the cost of accuracy or validity of transitive solutions. But lower accuracy or greater risk of biased conclusions may be an acceptable cost, and may even be beneficial if the situation would benefit more from an experience-based decision than a logic-based one.

Indeed, in one variant on the dual process theory (Reyna, 2004), it is argued that the more associative (gist-based) process often actually takes the lead in reasoning about risks, whether personal, medical or abstract. The argument is basically that rationality is not perfect deduction precisely because it involves these two processes both in parallel and in interaction (De Neys & Glumicic, in press). One implication is that in contrast with Evans' (2003) dual process theory, Reyna's (2004) theory holds that older more associative brain systems were not superseded by newer deductive ones, but the two work together to produce something that is more rational or adaptive than either on its own; sometimes driven more by one mode and sometimes more by the other (De Neys & Glumicic, in press; Kokis et al., 2002).

It is highly unfortunate that for over 3 decades now, the transitive research field has been

convinced that there is only one transitive mode. As argued by Wright (1998a), the insistence on the use of the B&T task has then led to the wrong (i.e., associative) mode to be taken as the only valid or useful (deductive) mode (Acuna et al., 2002; Breslow, 1981; Bryant, 1998; Bryant & Trabasso, 1971; Hummel & Holyoak, 2001; Siemann & Delius, 1998; Trabasso, 1977). Indeed, some are so committed to these views that they advocate manipulation of number of premises, amount of training, the form of feedback, or even the specific questions asked about comparison-pairs; all in order to guarantee the entire transitive series is represented in long-term memory before testing for transitivity (e.g., De Lillo et al., 2001; Holcomb et al., 1997; Russell et al., 1996; Stromer et al., 1993; Titone et al., 2004). For instance, Russell et al., trained 6 year-olds to 100% and then waited 12 minutes before testing, to ensure the entire series was in long-term memory. This is at odds with the dual process assertion about general reasoning, that the memory loads for deduction tend to be on working memory rather than on long-term memory, and on explicit representation rather than on implicit representation (Evans, 2003; Kokis et al., 2002). Indeed, data from Titone et al. (2004) intimate that procedures such as used by Russell et al. (1996) may invalidate the task regarding deduction-based transitive inference. As an alternative to Russell et al.'s approach, Acuna et al. (2002) increased the items in the series from 5 to 11 and gave additional training, and then requested answers only about distance from one end-anchor, to ensure a symbolic distance effect. We have already seen that this is unlikely to call for deduction, either for B&T tasks or our own task here.

But this does not mean that the associative mode has little use today. Allen (2006) points out that, particularly for non-humans, great importance may be attached to being able to place oneself along a transitive continuum such as the social rank of maybe 80 potential competitors. However, the importance of this kind of preoccupation is far lower for humans than for non-humans (Archie et al., 2006; De Lillo et al., 2001; Hummel & Holyoak, 2001; Siemann & Delius, 1998). For humans, concrete situations calling for a deductive transitive competence will more often involve fewer than five entities and will not include opportunities for a large number of repeated

exposures to premise information. Indeed, such situations will tend to centre around education, work and social problem solving; each of which is highly relevant to humans and largely irrelevant to non-humans (Allen, 2006; Goel et al., 2004; Markovits & Dumas, 1999; Russell et al., 1996).

Chapman (1999) argues that the target deductive mechanism in question must to some extent be founded in linguistic competencies (see also, Evans, 2003; Hummel & Holyoak, 2001; Sternberg, 1980; Reyna, 2004; Wright, 1998b), although Wright (2001) and Stromer et al. (1993) independently note that there is no current evidence that we need to impute a linguistic basis over and above a merely symbolic basis. Wright (2001) goes further, arguing that actually it may be the symbolic competence that underlies language acquisition in humans and also the transitive competencies of some, but almost certainly not all, non-human species (McGonigle & Chalmers, 1992).

To sum, we have presented theorising and findings that show that the generic B&T task is likely to have missed its target deductive transitive phenomenon, but this does not mean that it is irrelevant to improving our understanding of transitive inference. Also, transitive research still has plenty to offer current theories of wider reasoning. The challenge now is to modify a well specified model that already accommodates transitive reasoning (e.g., Bara et al., 2001; De Lillo et al., 2001; Halford & Andrews, 2004; Hummel & Holyoak, 2001; Wu & Levy, 2001) so that it now captures both the deductive and associative mechanisms for transitive inference. But there is so much more to be gained from better understanding transitive inference than that: The original task was intended to be discriminative, telling apart individuals/groups highly competent in making deductive transitive inferences from those lower in this competence (Piaget, 1970; Markovits et al., 1995). Whilst the B&T task generally shows itself to give positive results but is poor at distinguishing between groups, our alternative non-training single-response task offers to do both. Thus, it can contribute to issues such as which brain areas or functional neural systems play a primary role in deductive versus associative transitive inference, whether consciousness is

essential for deductive transitive inference, whether both forms of transitive inference are equally strongly related to language, whether the memory-independence effect applies to the deductive mode as well as associative mode, and which species are developing a deductive mode of transitive inference (Bara et al., 2001; Brainerd & Reyna, 1992; Goel et al., 2004; Martin & Alsop, 2004; Moses et al., 2006; Wright, 2001; Yamazaki, 2004). Of course it will definitively settle the original debate on the age children really become competent in deductive transitive inference (Holcomb et al., 1997; Wright, 2006a). But more applied-cognition questions can also now be tackled such as whether deduction is really diminished in schizophrenia, whether disabilities such as blindness have knock-on effects on deductive versus non-deductive transitive inference, whether spatial navigation is supported more by deductive or associative transitive inference, whether the deductive mode of transitive inference alters in old age, and whether associative training can improve deductive competencies in reasoners with learning difficulties (Ittyerah & Samarapungavan, 1989; Maydak et al., 1995; Stromer et al., 1993; Schnall & Gattis, 1998; Titone et al., 2004; Viskontas et al., 2005).

References

Acuna, B. D., Sanes, J. N., & Donoghue, J. P. (2002). Cognitive mechanisms of transitive inference. Experimental Brain Research, 146, 1-10.

Allen, C. (2006). Transitive inference in animals: reasoning or conditioned associations? In: S. Hurley, M. Nudds, (Eds.), Rational animals? (pp. 175–185). Oxford: Oxford University Press.

Archie, E. A., Morrison, T. A., Foley, C. A. H., Moss, C. J., & Alberts, S. C. (2006). Dominance rank relationships among wild female African elephants, Loxodonta Africana. Animal Behaviour, 71, 117-127.

Bara, B. G., Bucciarelli, M., & Lombardo, V. (2001). Model theory of deduction: A unified computational approach. Cognitive Science, 25, 839-901.

Bond, A. B., Kamil, A. C., & Bolda, R. P. (2003). Social complexity and transitive inference in Corvids. Animal Behaviour, 65, 479-487.

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. Developmental Review, 27, 41-74.

Brainerd, C. J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy-trace theory of transitivity development. Developmental Review, 4, 311-377.

Brainerd, C. J., & Reyna, V. F. (1992). Explaining "memory free" reasoning. Psychological Science, 3, 332-339.

Breslow, L. (1981). Re-evaluation of the literature on the development of transitive inferences. Psychological Bulletin, 88, 325-351.

Bryant, P. (1998). Cognitive Development. In Eysenck, M. Psychology an integrated approach. London: Wesley Longman Limited.

Bryant, P. E., & Trabasso, T. (1971). Transitive inferences and memory in young children. Nature, 232, 456-458.

Bryson, J. J., & Leong, C. S. (2007). Primate errors in transitive inference: a two tier learning

model. Animal Cognition, 10, 1-15

Chapman, M. (1999). Constructivism and the problem of reality. Journal of Applied Developmental Psychology, 20 (1), 31-43.

Dayton, C. M. (1998). Latent class scaling analysis. Thousand Oaks, CA: Sage.

De Lillo, C., Floreano, D., & Antinucci, F. (2001). Transitive choices by a simple, fully connected, backpropagation neural network: Implications for the comparative study of transitive inference. Animal Cognition, 4 (1), 61-68.

De Neys, W., & Glumicic, T. (in press). Conflict monitoring in dual process theories of thinking. Cognition.

Demarais, A. M., & Cohen, B. H. (1998). Evidence for image-scanning eye movements during transitive inference. Biological Psychology, 49, 229-247.

Eichenbaum, H. (2001). The hippocampus and declarative memory: Cognitive mechanisms and neural codes. Behavioural Brain Research, 127, 199-207.

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. Trends in Cognitive Sciences, 7 (10), 454-459.

Favrel, J., & Barrouillet, P. (2000). On the relation between representations constructed from text comprehension and transitive inference production. Journal of Experimental Psychology: Learning Memory and Cognition, 26 (1), 187-203.

Fernandez, G., & Tendolkar, I. (2001). Integrated brain activity in medial temporal and prefrontal areas predicts subsequent memory performance: Human declarative memory formation at the system level. British Research Bulletin, 55 (1), 1-9.

Ferreira, M. B., Garcia-Marques, L., Sherman, S. J., & Sherman, J. W. (2006). Automatic and controlled components of judgement and decision making. Journal of Personality and Social Psychology, 91 (5), 797-813.

Goel, V., Makale, M., & Grafman, J. (2004). The hippocampal system mediates logical reasoning about familiar spatial environments. Journal of Cognitive Neuroscience, 16 (4), 654-

662.

Goodwin, G. P., & Johnson-Laird, P. N. (2005). Reasoning about relations. Psychological Review, 112 (2), 468-493.

Halford, G. S., & Andrews, G. (2004). The development of deductive reasoning: How important is complexity. Thinking & Reasoning, 10 (2), 123-145.

Holcomb, W. L., Stromer, R., & Mackay, H. A. (1997). Transitivity and emergent sequence performances in young children. Journal of Experimental Child Psychology, 65, 96-104.

Hong, L., & Chond, L. (2001). The mental models of solving of three-term series problems by individuals. Acta Psychologica Sinica, 33 (6), 518-525.

Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. In M. Gattis (Ed.), Spatial schemas and abstract thought. (pp. 279-305). Cambridge MA: MIT Press.

Ittyerah, M., & Samarapungavan, A. (1989). The performance of congenitally blind children in cognitive developmental tasks. British Journal of Developmental Psychology, 7, 129-189.

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: age trends and associations with cognitive ability and cognitive styles. Journal of Experimental Child Psychology, 93, 26-52.

Lazareva, O. F., & Wasserman, E. A. (2006). Effect of stimulus orderability and reinforcement history on transitive responding in pigeons. Behavioural Processes, 72, 161-172.

Markovits, H., & Dumas, C. (1992). Can pigeons really make transitive inference? Journal of Experimental psychology, Animal Behaviour Processes, 18, 311-312.

Markovits, H., & Dumas, C. (1999). Developmental patterns in the understanding of social and physical transitivity. Journal of Experimental Child Psychology, 73, 95-114.

Markovits, H., Dumas, C., & Malfait, N. (1995). Understanding transitivity of a spatial relationship, A development analysis. Journal of Experimental Child Psychology, 59, 124-141.

Martin, N., & Alsop, B. (2004). Transitive inference and awareness in humans. Behavioural Processes, 67, 157-165.

Maydak, M., Stromer, R., Mackay, H. A., & Stoddard, L. T. (1995). Stimulus classes in matching to sample and sequence production, the emergence of numeric relations. Research in Developmental Disabilities, 16 (3), 179-204.

McGonigle, B. O., & Chalmers, M. (1992). Monkeys are rational. Quarterly Journal of Experimental Psychology, 45B, 189-228.

Moses, S. N., Villate, C., & Ryan, J. D. (2006). An investigation of learning strategy supporting transitive inference performance in humans compared to other species. Neuropsychologia, 44 (8), 1370-1387.

Murayama, T., & Tobayama, T. (1997). Preliminary study on stimulus equivalence in Beluga (Delphinapterus leucas) Japanese Journal of Animal Psychology, 47**,** 79–89.

Perner, J., & Aebi, J. (1985). Feedback-dependent encoding of length series. British Journal of Developmental Psychology, 3, 133-141.

Perner, J., & Mansbridge, D. G. (1983). Developmental differences in encoding length series. Child Development, 54, 710-719.

Piaget, J. (1970). Piaget's theory. In P. H. Mussen (Ed.), Carmichael's manual of child psychology. New York: Wiley.

Reyna, V. F. (2004). How people make decisions that involve risk. A dual-processes approach. Current Directions in Psychology Science, 13 (2), 60-66.

Russell, J., McCormack, T., Robinson, J., & Lillis, G. (1996). Logical (versus associative) performance on transitive reasoning tasks by children: Implications for the status of animals' performance. The Quarterly Journal of Experimental Psychology, 49B, (3), 231-244.

Schnall, S., & Gattis, M. (1998). Transitive inference by visual reasoning. In M. A. Gernsbacher, & S. J. Derry (Eds.), Proceedings of the twentieth annual conference of the cognitive science society. (pp. 929-934). Mahwah, NJ: Erlbaum.

Shafir, S., Waite, T. A., & Smith, B. H. (2002). Context-dependent violations of rational choice in honeybees (Apis mellifera) and grey rays (Perisoreus Canadensis). Behav. Bcol. Sociobiol., 51, 180-

187.

Siemann, M., & Delius, J. D. (1996). Influences of task concreteness upon transitive responding in humans. Psychological Research, 59, 81-93.

Siemann, M., & Delius, J. D. (1998). Algebraic learning and neural network models for transitive and non-transitive responding. European Journal of Cognitive Psychology, 10 (3), 307-334.

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. Journal of Experimental Psychology: General, 109 (2), 119-159.

Stromer, R., Mackay, H. A., Cohen, M., & Stoddard, L. T. (1993). Sequence learning in individuals with behavioural limitations. Journal of Intellectual Disability Research, 37, 243-261.

Titone, D., Ditman, T., Holzman, P. S., Eichenbaum, H., & Levy, D. L. (2004). Transitive inference in Schizophrenia: Impairments in relational memory organization. Schizophrenia Research, 68, 235-247.

Trabasso, T. (1977). The role of memory as a system in making transitive inferences. In R. V. Kail & J. W. Hagan (Eds.), Perspectives on the Development of Memory and Cognition. (pp. 333-366). Hillsdale NJ: Erlbaum.

Viskontas, I. V., Holyoak, K. J., & Knowlton, B. J. (2005). Relational integration in older adults. Thinking & Reasoning, 11 (4), 390-410.

Wright, B. C. (1998a). And if the developmental data doesn't quite fit... Behavioral and Brain Sciences, 21, 847-848.

Wright, B. C. (1998b). Psychological mechanisms of logical transitive inference in adults and children. Doctoral Dissertation: University of Oxford, England.

Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. Developmental Review, 21, 375-422.

Wright, B. C. (2006a). On the emergence of the discriminative mode for transitive-inference. European Journal of Cognitive Psychology, 18 (5), 776-800.

Wright, B. C. (2006b). The information processing task revisited: Investigating profiles from the start to the end of training. Thinking & Reasoning, 12 (1), 91-123.

Wright, B. C., & Dowker, A. D. (2002). The role of cues to differential absolute size in children's transitive inference. Journal of Experimental Child Psychology, 81, 249-275.

Wright, B. & Wanley, A. (2003). Adults' versus children's performance on the Stroop task: Interference and facilitation. British Journal of Psychology, 94, 475-485.

Wu, X., & Levy, W. B. (2001). Simulating symbolic distance effects in the transitive inference problem. Neurocomputing, 38-40, 1603-1610.

Wynne, C. D. L. (1998). A minimal model of transitive inference. In C. D. L. Wynne & J. E. R. Staddon (Eds.), Models for action: Mechanisms for adaptive behavior. Hillsdale NJ: Erlbaum.

Yamazaki, Y. (2004). Logical and illogical behavior in animals. Japanese Psychological Research, 46 (3), 195-206.

Table 1: Summary of Correct Responses

	Comparison-Pairs				Overall
0-Step	A:B 49.091	B:C 64.848	C:D 58.636	D:E 87.424	65.000
	(4.677)	(2.395)	(2.475)	(1.512)	(2.036)
1-Step		A?C 37.576	B?D 55.152	C?E 82.273	58.333
		(3.624)	(2.157)	(1.935)	(1.645)
2-Step			A?D 38.182	D?E 84.545	61.364
			(3.781)	(2.210)	(1.649)
3-Step				A?E 79.848	79.849
				(1.856)	(1.856)
Overall				63.758	
				(1.524)	

Note, Numbers in parentheses are Standard Errors

Table 2: Summary for the Two Response-Time Measures

	Comparison Pairs nRT				Overall
0-Step	A:B 2502 (88)	B:C 2871 (119)	C:D 2843 (90)	D:E 2700 (113)	2729 (89)
1-Step		A?C 2942 (112)	B?D 3131 (137)	C?E 2965 (148)	3013 (117)
2-Step			A?D 2837 (125)	B?E 2715 (104)	2776 (105)
3-Step				A?E 2737 (116)	2737 (116)
Overall				2825 (95)	
	Comparison Pairs cRT				
0-Step	A:B 2784 (111)	B:C 2922 (123)	C:D 3142 (3142)	D:E 2714 (111)	2891 (104)
1-Step		A?C 3573 (166)	B?D 3606 (177)	C?E 3131 (170)	3437 (156)
2-Step			A?D 3388 (169)	B?E 3001 (167)	3195 (153)
3-Step				A?E 2969 (142)	2969 (142)
Overall				3123 (119)	

Note, Numbers in parentheses are Standard Errors

Table: 3: Summary and Evaluation Regarding B&T Phenomena

	Response-Accuracy	Fits B&T	nRT	Fits B&T	cRT	Fits B&T
Lexical Marking:	56.97 (3.226)	NO	2687 (98)	NO	2853 (106)	NO
Unmarked Premises						
Marked Premises	73.03 (1.540)		2771 (91)		2928 (111)	
Inward-Acquisition:	68.258 (2.782)	YES	2601(94)	YES	2749 (100)	YES
Outer Premises						
Inner Premises	61.742 (2.628)		2857 (94)		3032 (115)	
1-Step: A?C Inference	37.576 (3.624)	NO	2942 (112)	NO	3573 (166)	NO
A?C Mean Antecedents	56.969 (3.536)		2687 (103)		2853 (117)	
1-Step: *B?D Inference	55.152 (2.157)	NO	3131 (137)	NO	3606 (177)	NO
B?D Mean Antecedents	61.742 (2.435)		2857 (104)		3032 (125)	
1-Step: C?E Inference	82.273 (1.935)	YES	2965 (148)	NO	3130 (170)	NO
C?E Mean Antecedents	73.03 (1.994)		2772 (101)		2928 (119)	
Serial Position	NOT A OR E	NO	NOT A OR E	NO	NOT A OR E	NO
End-Anchor	REVERSED	NO	ABOLISHED	NO	REVERSED	NO
Symbolic Distance	NON STANDARD	NO	ABOLISHED	NO	NON STANDARD	NO
Overall Series	LOW	NO	SLOW	NO	SLOWER	NO

Note, Numbers in parentheses are Standard Errors

Figure Captions

Figure 1 Top – Response-accuracy for the end-item effects. The E end-anchor was consistently superior to the A end-anchor. Middle – For nRT the difference between end-anchors favoured item-A for 0-step pairs but favoured item-E for 2-step pairs. Bottom – For cRT, pairs involving item-E were consistently responded to faster than those involving item-A.

Figure 2 Top – Response-accuracy by serial position. Item-A did not fit the serial position effect, and nor did the gradation from item-B to item-D. Bottom - For non-discriminative-RT (nRT), performance slowed from the end towards the centre; but item-E was faster than item-A. For correct-only-RT (cRT), there was no systematic pattern of serial position.

Figure 1

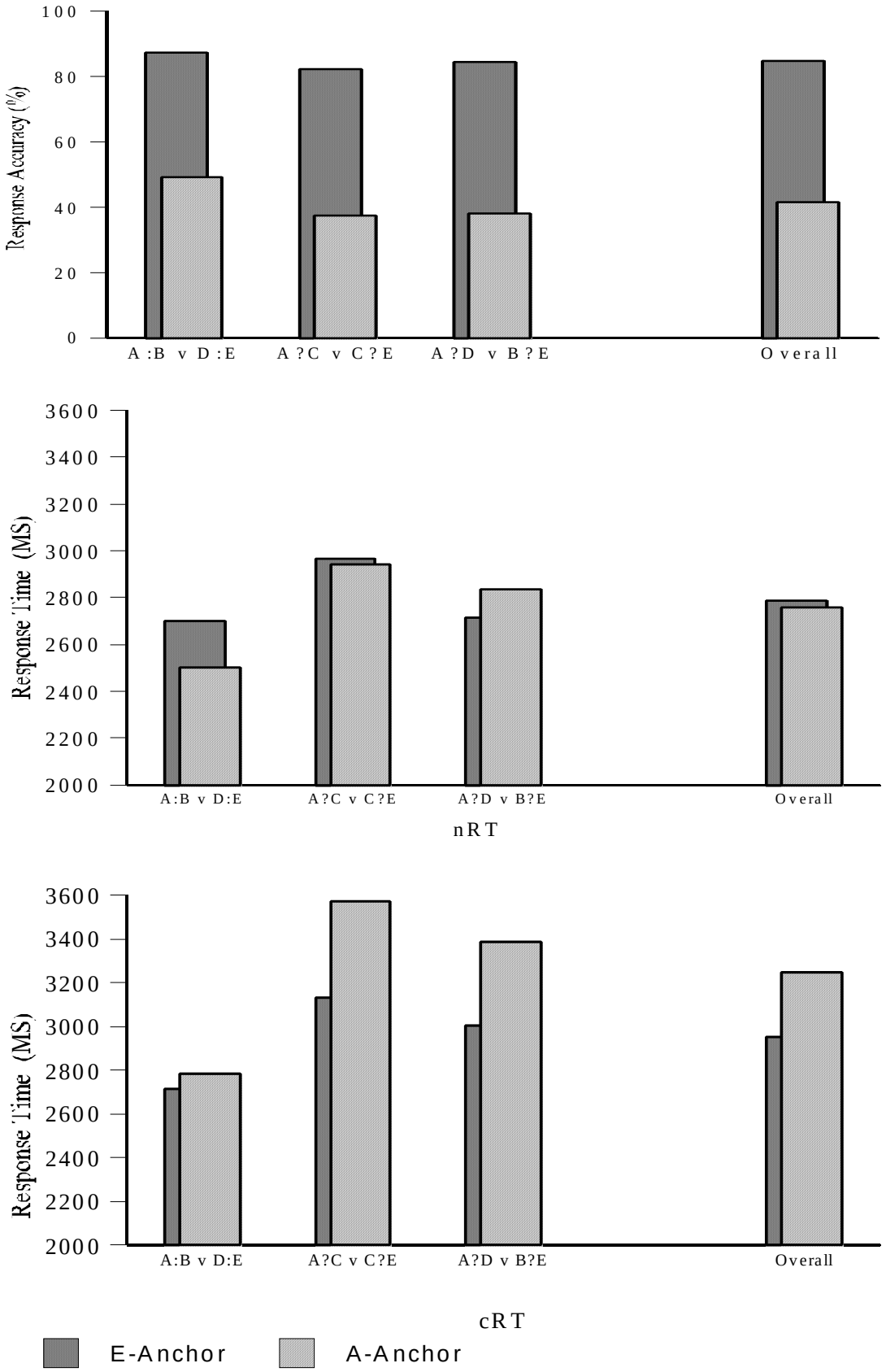
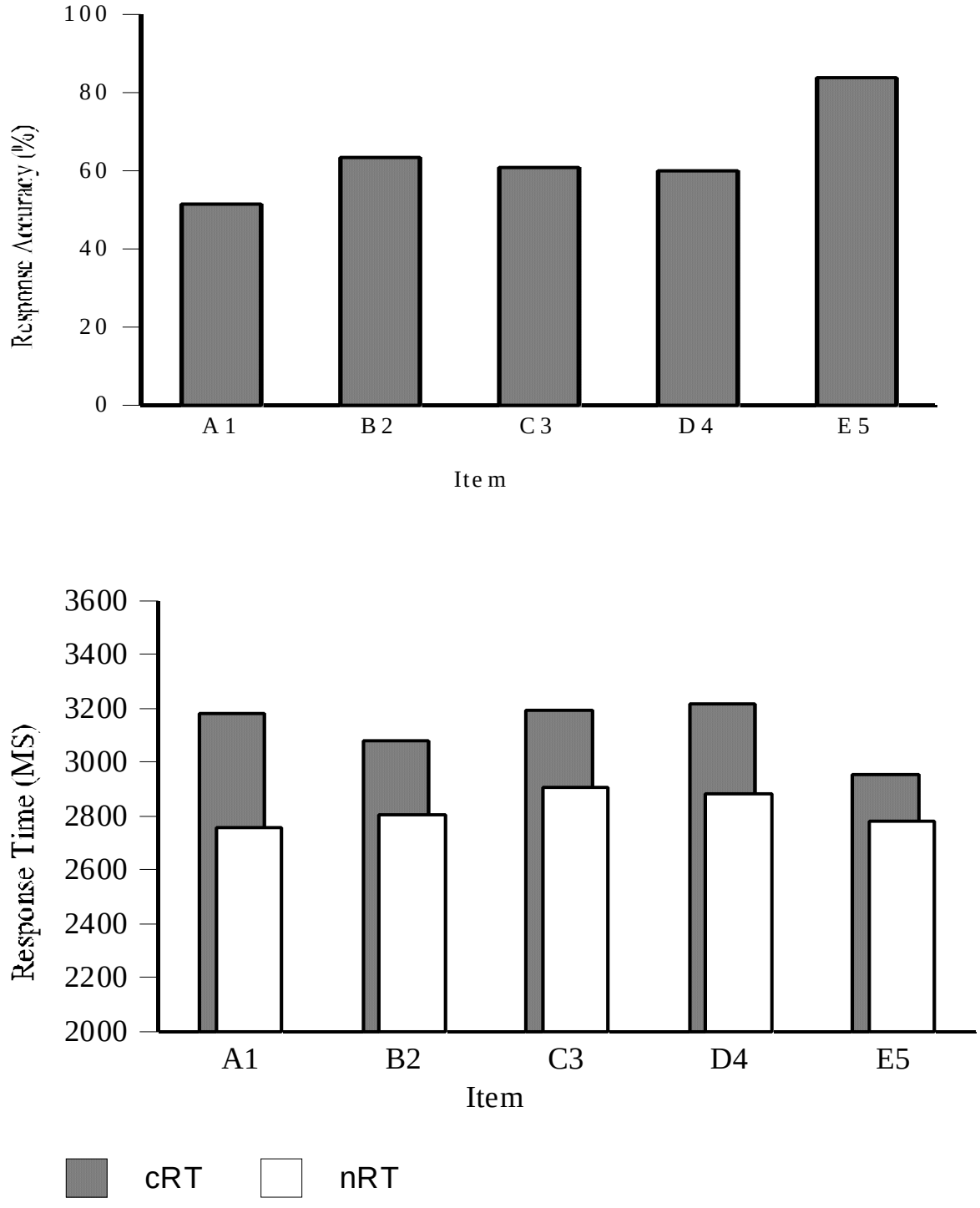

Response Accuracy (%)
100
80
60
40
20
0
A:B v D:E
A?C v C?E
A?D v B?E
Overall
Response Time (MS)
3600
3400
3200
3000
2800
2600
2400
2200
2000
A:B v D:E
A?C v C?E
A?D v B?E
Overall
nRT
Response Time (MS)
3600
3400
3200
3000
2800
2600
2400
2200
2000
A:B v D:E
A?C v C?E
A?D v B?E
Overall
cRT
E-Anchor
A-Anchor

Figure 2


100
80
60
40
20
0
Response Accuracy (%)
A 1
B 2
C 3
D 4
E 5
Item
3600
3400
3200
3000
2800
2600
2400
2200
2000
Response Time (MS)
A1
B2
C3
D4
E5
Item
cRT
nRT