Equal Treatment for all Antecedents: How Children Succeed with Principle B

Anastasia Conroy\textsuperscript{1}, Eri Takahashi\textsuperscript{1}  
Jeff Lidz\textsuperscript{1,2}, Colin Phillips\textsuperscript{1,2}

\textsuperscript{1} Department of Linguistics, \textsuperscript{2} Neuroscience and Cognitive Science Program  
University of Maryland, College Park

\textbf{Abstract.} A long-standing finding in the acquisition of anaphora is that children behave poorly with respect to Principle B, accepting locally bound pronouns until age 5 or later. A theoretically influential finding is that this Delay of Principle B Effect (DPBE) is not uniform across antecedents: children show a Quantificational Asymmetry (QA), demonstrating mastery of Principle B with quantified antecedents before referential antecedents (e.g., Chien \& Wexler 1990, Thornton \& Wexler 1999). This finding has been interpreted as dramatic evidence for syntactic theories that restrict the scope of binding constraints to bound variable anaphora (e.g., Reinhart 1983). However, the QA has been challenged, based upon discrepant findings and methodological concerns. Furthermore, the DPBE is puzzling in light of young children’s success with another disjoint reference rule, Principle C. Here we attempt to resolve the status of the QA and DPBE, based on 3 Truth Value Judgment Task studies and a meta-analysis of over 30 previous studies. Using improved experimental tests, we show that children disallow local pronoun binding with referential and quantificational antecedents when Principle B is at issue (Experiment 1), but freely allow local pronoun binding in identical scenarios when Principle B is neutralized (Experiment 2). When methodological flaws are reintroduced we replicate the finding of a QA (Experiment 3). These findings vindicate a claim by Elbourne (2005) that the QA is a methodological artifact, but disconfirm Elbourne’s prediction of an across-the-board DPBE. A meta-analysis of previous studies supports our conclusions: many studies that report a QA contain similar methodological errors to the ones we investigate, and the best evidence for DPBE indicates that it is a weaker effect than is commonly supposed. We suggest that the residual contrast in children’s ability to use Principles B and C to block ungrammatical antecedents mirrors a contrast found in adults in the effects of these constraints on real-time reference resolution.

1. Introduction

It is rare that data from child language are taken to distinguish competing theoretical positions. One such case concerns what Elbourne (2005) calls the \textit{Quantificational Asymmetry} (QA) in children’s application of Principle B, a constraint that bars local antecedents for pronouns. Many studies, spanning the past 25 years, have reported that 4-6 year old children allow a non-adult interpretation of (1) that is equivalent to \textit{Mama Bear washed herself}. However, when the referential subject in (1) is replaced with a quantificational subject, as in (2), the corresponding anaphoric interpretation is no longer allowed by children.

(1) Mama bear washed her.  
(2) Every bear washed her.

The theoretical argument, due originally to Chien \& Wexler (1990), concerns the scope of Principle B in the grammar. Whereas the classical binding theory of Chomsky (1981) treats all cases of anaphora alike as involving coin dexation, Reinhart’s (1983) approach distinguishes bound variable anaphora from other cases of coreference. Consequently, the quantificational asymmetry in children’s judgments is taken to mirror an asymmetry in the structure of Reinhart’s binding theory, and has been widely understood to provide strong evidence for theories that contain that asymmetry.
However, the evidence for the Delay of Principle B Effect (DPBE) in children creates a further asymmetry that fits less easily with theoretical accounts of anaphora. Children as young as 3 years of age show mastery of Principle C, a constraint on backwards anaphora that rules out coreference in sentences like (3) (Crain & McKee 1985, Crain & Thornton 1998, Guasti & Chierchia 1999/2000, Kazanina & Phillips 2001; Leddon & Lidz 2005). The developmental advantage for Principle C over Principle B is unexpected under most theoretical accounts.

(3) She washed Mama Bear.

In this article we consider children’s knowledge of Principle B in detail and suggest that many of the prior observations reflect shortcomings of the experimental tests used and not properties of children’s grammars. Related concerns about the evidence for the QA have been presented by Elbourne (2005), although without new developmental evidence. Results from three new experiments partly vindicate Elbourne’s concerns, in the respect that appropriately controlled experiments appear to eliminate the QA. However, our findings disconfirm Elbourne’s prediction that children would show a strong across-the-board DPBE. On the contrary, we find that 4-year old children overwhelmingly respect Principle B in sentences with referential and quantificational antecedents alike, although our analysis of previous studies suggests that children’s tendency to assent to interpretations that violate Principle B is genuine, and substantially modulated by the details of experimental designs. We conclude that there is no QA, and that at the level of grammatical representation there is no DPBE. However, we also argue that children are more susceptible to errors that violate Principle B than they are to errors that violate Principle C, and that this contrast derives from independently motivated differences found in the processing of Principles B and C in adults. Our studies therefore remove one of the arguments in favor of Reinhart’s theory of binding, although we remain neutral as to whether this theory is actually correct. At the same time, our studies remove a potential impediment to theories that give parallel accounts of Principles B and C.

In Section 2 we summarize two lines of evidence for the distinction between bound variable anaphora and coreference, drawing on classic theoretical and developmental arguments. We also highlight the contrasting experimental findings about children’s mastery of Principles B and C. In Section 3 we review the methodological assumptions behind widely used tests of children’s grammars generally, and Principle B in particular. In Section 4 we present 3 new experiments that address the methodological concerns, showing no QA and no DPBE, and leaving us with the question of why findings about children’s knowledge of Principle B are so varied. In Section 5 we review over 30 previous studies on binding constraints in children, and relate these findings to recent studies of binding constraints in real-time language processing in adults, concluding that although children’s grammars are apparently intact, they show exaggerated susceptibility to illicit antecedents that are also fleetingly considered in studies with adults.

2. **Asymmetries in Binding Theory**

2.1 **Two Types of Anaphoric Relation**

At bottom, the binding theoretic debate centers around the formal mechanisms underlying anaphora in the grammar and begins with the discovery that sentences like (4) are three-ways ambiguous, not two.
(4) Al loves his sister.

One reading is the one in which the pronoun refers to a sentence external antecedent. However, even if we restrict the pronoun to be anaphoric to the subject of the sentence, we still find an ambiguity. The ambiguity can be seen more clearly when we place such sentences in VP-ellipsis contexts, as in (5) (Sag 1976, Williams 1977).

(5) Al loves his sister and Bill does too
   a. \( \alpha \) = Al\(_i\) loves his\(_i\) sister and Bill\(_i\) loves his\(_j\) sister too (‘sloppy’ reading)
   b. \( \beta \) = Al\(_i\) loves his\(_i\) sister and Bill\(_j\) loves his\(_i\) sister too (‘strict’ reading)

We can interpret the second conjunct as meaning that Bill loves Bill’s sister or as meaning that Bill loves Al’s sister. What this simple paradigm tells us is that there is more than one way for a pronoun to be connected to its antecedent. On the one hand, the pronoun may be treated as a bound variable, whose reference is determined by its antecedent. This gives rise to the so-called ‘sloppy’ interpretation (5a), because the elided VP and the overt VP each contain a bound variable pronoun bound by the subject of the corresponding clause. Alternatively, the pronoun may be understood to have fixed reference that happens to match the reference of the subject of the first clause. This yields the so-called ‘strict’ interpretation (5b). In this case the elided VP, like its overt counterpart, contains a pronoun that refers to the subject of the first clause. The pronoun in the first clause corefers with the subject of the clause, but unlike a bound variable it is not directly dependent on the subject for its reference. This type of coreference is sometimes referred to as ‘accidental coreference’.

Further evidence for the ambiguity between bound variable anaphora and coreference comes from cases in which a pronoun takes a quantificational antecedent, as in (6).

(6) Every linguist loves his sister and Bill does too
   a. \( \alpha \) = every linguist\(_i\) loves his\(_i\) sister and Bill\(_j\) loves his\(_j\) sister too (sloppy)
   b. \( \beta \) \neq \text{every linguist}\(_i\) loves his\(_i\) sister and Bill\(_j\) loves his\(_j\) sister too (strict)

Because quantifiers are not referential, no accidental coreference is possible. Consequently, no strict reading is possible in the VP ellipsis context (6b). The overt pronoun can only be connected to a quantificational antecedent as a bound variable, and consequently the elided VP must also contain a bound variable (6a). In sum, data of the kind in (5-6) show that there are two mechanisms by which a pronoun may be linked to its antecedent: one involving variable binding and the other involving coreference (Sag 1976, Williams 1977, Evans 1980, Higginbotham 1983).

Reinhart (1983) observed further that the syntactic conditions on bound variable anaphora are stricter than those on accidental coreference. For example, a bound variable must be c-commanded by its antecedent (7a), whereas no such restriction holds for accidental coreference (7b).

(7) a. The people who work for him\(_{ij}\) love every department chair\(_i\).
   b. The people who work for him\(_{ij}\) love Al\(_i\).
Similarly, in VP-ellipsis contexts a pronoun with a non c-commanding antecedent supports only a coreferential reading, and hence only allows the strict reading of the elided VP (8).

(8) The people who work for Al love him and the people who work for Bill do too.
   a. = ... and the people who work for Bill love Al.
   b. ≠ ... and the people who work for Bill love Bill.

Like variable binding in general, the binding conditions (Chomsky 1981, 1986) are stated in terms of the c-command relation Principle B rules out a pronoun with a local antecedent only if c-command obtains between them (9). Similarly, Principle C rules out an R-expression (i.e., non-pronominal NP) that is c-commanded by a coreferential NP (10).

(9) a. Al\textsubscript{i} likes him\textsubscript{ij}
   b. Al\textsubscript{i}'s sister likes him\textsubscript{ij}

(10) a. He\textsubscript{ij} thinks that I like Al\textsubscript{i}
    b. His\textsubscript{ij} sister thinks that I like Al\textsubscript{i}

2.2 The Scope of Binding Constraints

Reinhart (1983) argued that the theory of binding should apply only to bound variable anaphora, and not to anaphoric interpretations in general. We summarize here one theoretical consideration and an empirical argument that is particularly relevant to the developmental issues that are our main interest here.

A primary theoretical motivation for Reinhart’s approach comes from the fact that the disjoint reference constraints (i.e., Principles B and C) and bound variable anaphora apply only to c-command relations. This parallel is expected if the binding constraints are restricted to bound variable anaphora.

This assumption straightforwardly predicts that the bound variable interpretation of the pronoun in (11) is blocked by Principle B. However, the prediction is more complicated for sentences like (12) that have a referential subject NP.

(11) Every candidate, likes him\textsubscript{ij}.
(12) Al\textsubscript{i} likes him\textsubscript{ij}.

Recall that a referential NP may serve as the antecedent of a pronoun either via variable binding or via ‘accidental coreference’. Therefore, if accidental coreference is generally available and if Principle B regulates only the distribution of bound variables, then it follows that the accidental coreference interpretation should be available in (12), yielding the interpretation that Al likes himself, contrary to speaker judgments. Reinhart argues that this prediction is not a shortcoming of her theory, but rather is a virtue. She claims that the accidental coreference representation is indeed available, but that special discourse circumstances are required to realize this possibility (Evans 1980, Higginbotham 1983). In a sentence like (13), for example, it is argued to be natural to interpret the pronoun him as coreferential with Bill, despite the fact that it is locally c-commanded by its antecedent in the last conjunct. Therefore, it is important to explain what distinguishes (12) from (13).
(13) I know what Mary, Sue and Bill have in common. Mary likes him, Sue likes him, and Bill likes him too.

Reinhart recognizes that her version of Principle B is insufficient to block coreference in (12), due to the possibility of accidental coreference.\footnote{Indeed, it is precisely this issue that led to Lasnik’s (1976) argument that the constraints on pronominalization must be stated in terms of disjoint reference and not coreference. However, Evans (1980) argued that even this restriction was insufficient to block accidental coreference. For alternative accounts of how to distinguish the coreference possibilities in (12) and (13), see Heim (1998) and Levinson (2000).} She proposes that this binding theoretic loophole is closed by an additional constraint, labeled Rule I (14).

(14) \textit{Rule I: Intrasentential Coreference} (Grodzinsky & Reinhart 1993, p. 79)

NP A cannot corefer with NP B if replacing A with C, C a variable A-bound by B, yields an indistinguishable interpretation.

In essence Rule I is an economy condition, stating that accidental coreference is possible only when bound variable anaphora is not. This successfully blocks accidental coreference in (12), because the bound variable interpretation and the accidental coreference interpretation have identical truth conditions. Furthermore, Rule I also provides an account of why coreference is possible in sentences like (13). The relevant interpretation of (13) asserts that that the property shared by Mary, Sue and Bill is the property of liking Bill, i.e., (15).

(15) \[ \lambda x. \ x \text{ likes Bill} \]

However, if the pronoun in the third conjunct is taken to be a bound variable, then that clause asserts that Bill is a self-liker, i.e., (16). It is clear that liking oneself and liking Bill are different properties for the members of a group to share, and therefore Rule I does not apply. Consequently, accidental coreference is not trumped by bound variable anaphora in (13) and hence the third conjunct is predicted to allow the interpretation that Bill likes himself.

(16) \[ \lambda x. \ x \text{ likes } x \]

The critical conclusion from this line of argument is that cases of coreference failure like (12), which under classic binding theory (Chomsky 1981) were taken to be violations of Principle B alone, are considered to be ruled out by two separate principles of grammar: Principle B (a condition on the syntax-semantics interface) and Rule I (a condition on the semantics-pragmatics interface).

2.3 \textit{A Developmental Dissociation, and a Further Asymmetry}

While the theoretical arguments in favor of restricting Principle B to cases of bound variable anaphora can stand on their own, an additional source of data is often lauded as the best possible evidence for the existence of two principles related to Principle B effects (Chien and Wexler 1990, Grodzinsky and Reinhart 1993, Thornton and Wexler 1999). This evidence comes from an apparent dissociation in children’s interpretation of sentences traditionally captured by Principle
B. Many studies of English and other languages, most notably Dutch and Russian, have reported that children incorrectly allow local binding of a pronoun with a referential antecedent until roughly age 5, but a number of these studies have reported no such delay in sentences with quantificational antecedents (see Section 5 for extensive references). For example, in a classic picture matching study (Chien & Wexler 1990) children in the 5-year old age group accepted coreference for sentences like *Mama Bear is touching her* on 51% of trials, but for sentences like *Every bear is touching her* on only 16% of trials.

This developmental dissociation, which Elbourne (2005) calls the Quantificational Asymmetry (QA) appears to provide striking support for Reinhart’s account of binding theory. The theory predicts that different mechanisms restrict quantificational and referential antecedents for pronouns, and the child data comports well with this prediction. If the children have already mastered Principle B, but either do not know or cannot apply the additional constraint that blocks local anaphora with referential antecedents, then the QA is captured.\(^2\)

However, there are a number of concerns about the strength of the theoretical conclusions that can be drawn from reports of the QA. First, although the QA has become a part of the received wisdom about language acquisition (e.g., Guasti 2004), the empirical record is not unequivocal (for reviews see Kaufman 1994, Koster 1994, Elbourne 2005). A number of studies have obtained discrepant findings, both regarding the QA and the strength of the DPBE. For example, studies of DPBE have found rates of acceptance of Principle B violations that range from 16-82%, i.e., far greater variability than would be expected by chance. We discuss the previous literature in more detail in Section 5.

Second, examination of the studies that have been used to show the QA raises concerns about the adequacy of the experimental designs used in these studies. Some of these concerns were already raised by Elbourne (2005), and we raise a number of additional concerns in Section 3.

Third, the developmental findings do not comport so well with Reinhart’s theory of binding as is sometimes suggested. It should be recalled that the theoretical arguments that we reviewed for Principle B can also be constructed for Principle C. Indeed, on Reinhart’s theory Principle C effects are governed strictly by Rule I. Thus, we would expect that children’s failure to apply Rule I, giving rise to apparent Principle B violations, should also give rise to apparent Principle C violations. But this appears not to be the case. Instead, a number of studies have found that Principle C is uniformly obeyed by children at age 4 and even younger (Crain & McKee 1985, Crain & Thornton 1998, Guasti & Chierchia 1999/2000, Kazanina & Phillips 2001, Leddon & Lidz 2005). Thus, to the extent that the Quantificational Asymmetry provides evidence for the necessity of a theoretical construct like Rule I, children’s success with Principle C remains a mystery. Put differently, to the extent that the developmental evidence shows an asymmetry between Principle B and Principle C, it casts doubt upon the theoretical claim that Principles B and C are governed by common mechanisms. This concern applies not just to Reinhart’s theory of binding, but to any approach that treats the disjoint reference rules similarly.

With these issues in mind, the next section aims to identify the features of a fair test of the QA and the DPBE, using as a focus the Truth Value Judgment Task (TVJT; Gordon 1996, Crain

\(^2\) The literature contains a number of different accounts of the specific cause of the DPBE. For example, Reinhart has argued that children have full adult knowledge, but are unable to perform the computations needed to implement Rule I (Grodzinsky & Reinhart 1993, Reinhart 2006), whereas Wexler and his colleagues have argued that children lack adult knowledge of a pragmatic principle (‘Principle P’) that restricts accidental coreference (Chien & Wexler 1990, Thornton & Wexler 1999). The difference between these accounts does not affect our arguments in this paper.
& Thornton 1998), although many of the issues raised generalize to other experimental measures.

3. Truth Value Judgment Tests

3.1 The Logic of the Task

Much of the evidence for the Delay of Principle B Effect, and for the Quantificational Asymmetry in particular, has been drawn from studies using the Truth Value Judgment Task (TVJT), a task that confers many advantages but that requires great care in its use and interpretation. In this section, we briefly summarize the desiderata for a fair test of children’s knowledge of constraints on binding and coreference, and raise concerns about how well these desiderata have been satisfied in previous studies, expanding upon the critique of Elbourne (2005). Our discussion here focuses on the TVJT, but similar considerations apply to other experimental tasks that have been used to test children’s knowledge of binding constraints.

Suppose that we want to know what interpretations 4-year old children allow for sentences like (17) and (18), which have been the focus of most research on the DPBE and the QA. Our interest is in whether children respect the constraint that prevents the pronoun him from being anaphoric to a local subject NP in the adult grammar, and whether this constraint impacts the two examples equally. In what follows, we use the term anaphoric as a cover term for variable binding and (accidental) coreference, reserving the terms bound and coreferential for the two specific types of referential dependencies. We refer to interpretations in which the pronoun lacks an intrasentential antecedent as deictic. We also describe cases of illicit coreference and cases of illicit bound variable anaphora as Principle B violations, with no intended prejudice on the question of whether these cases should be handled by one vs. two mechanisms in the grammar.

(17) Grumpy painted him.
(18) Every dwarf painted him.

Clearly, we cannot ask young children to give us explicit judgments about the range of coreference possibilities for such sentences. The Truth Value Judgment Task was devised to tackle the problem of probing complex grammatical phenomena in young children using a task that is engaging and requires simple judgments from children. (Gordon 1996, Crain & Thornton 1998). In a TVJT a child and a puppet, such as Kermit the Frog, together watch an experimenter tell a story with props. Following the story, Kermit makes a statement about the story and the child’s task is to judge whether Kermit’s statement was accurate. The task has many advantages. It can be used to probe grammatically sophisticated issues with young children (most children aged 4-years and up, and some 3-year olds). It is engaging and non-confrontational, and it has special advantages for testing sentences that have multiple potential interpretations. Another advantage of the task is that the test sentence is presented in the context of a discourse, thereby encouraging the experimenter to make explicit assumptions about the impact of discourse context on sentence interpretations. When it is used as a test of binding constraints, the core logic of the TVJT is straightforward: if a child encounters a story in which the anaphoric interpretation of the pronoun in (17) or (18) is both true and prominent in the story, but does not judge (17) or (18) to be true statements about the story, then the child presumably did not access the anaphoric interpretation of the pronoun, possibly because a binding constraint made that interpretation
inaccessible. However, the simple judgment that the TVJT requires of a child also presents its greatest challenges. Great care must be taken in order to satisfy the assumptions underlying the task, in order to avoid misleading results, as Crain and Thornton and their colleagues have argued extensively (e.g., Crain & Thornton 1998).

One of the fundamental assumptions underlying the TVJT is that children will assent to the truth of a sentence if they can. In the case of potentially ambiguous sentences, sentences are presented in scenarios that make only one reading true at a time. Thus, if a child assents to the truth of the sentence in a scenario, we know that the child has access to the interpretation made true in that scenario. If the child rejects the sentence in a scenario, then that is taken as evidence that the interpretation made true in that scenario is unavailable. Of course, this inference is only valid if the rejected interpretation was, in fact, made available by the scenario. In other words, the experimenter must be sure that all interpretations under investigation are made available in the experimental contexts. Otherwise, a child may reject a sentence not because the true interpretation is absent from his grammar, but because that interpretation is not under consideration in the context.

Following this logic, TVJT tests of binding constraints rely on the assumption that the test scenario makes two different interpretations accessible (though not necessarily true): one interpretation corresponds to the anaphoric interpretation of the pronoun, and the other interpretation corresponds to the deictic interpretation of the pronoun. Children then judge the truth of a potentially ambiguous sentence, and the researcher can use the children’s judgments to infer which interpretation of the sentence was accessed. If children’s judgments show that they systematically fail to access one interpretation, then the researcher may conclude that the children do not allow that interpretation of the test sentence.

However, it is important to recognize that when a child rejects a sentence like (17) or (18) in a TVJT scenario where the anaphoric interpretation is true, the child is not a ‘little linguist’ who is directly judging the anaphoric interpretation as ungrammatical. Rather, it is assumed that the child is denying the truth of an alternative, deictic interpretation of the pronoun. When a child does this, it is commonly assumed that the child chose to focus on the deictic interpretation because the anaphoric interpretation was unavailable, and furthermore that it was the child’s grammar that made the anaphoric interpretation unavailable, and not some other factor. However, this depends on the further assumption that the child would have assented to the anaphoric interpretation, if only his grammar had not blocked that interpretation. For these reasons, when children reject target sentences in TVJT tests of binding constraints, the rejections provide only indirect evidence that the relevant interpretations are ungrammatical for the children.

Let us now consider how the assumptions of the TVJT must be satisfied in the context of a Principle B experiment, and the implications for specific experimental designs. At the most general level, these assumptions boil down to the need to balance the relative availability of the interpretations under investigation.

---

3 Crain and Thornton (1998) correctly point out that children’s negative judgments in a TVJT provide stronger evidence for the effects of a grammatical constraint if the story is presented in a way that is biased towards consideration of the ungrammatical-but-true interpretation. However, this benefit obtains only if the child’s interpretation contradicts the bias in the story. If a story strongly biases an ungrammatical interpretation and the child assents to that interpretation, it is impossible to know how much this reflects a non-adultlike grammar or experimental bias.
Assumption 1: Anaphoric Interpretation. The anaphoric interpretation is readily accessible and felicitous in the test scenario; the grammar is the only mechanism that could cause the anaphoric interpretation of the test sentence to be inaccessible. If a child has any reason independent of Principle B to ignore the anaphoric interpretation of the pronoun in (17) and (18), then the logic of the TVJT fails and the task does not test knowledge of Principle B. For example, if a child avoids the anaphoric interpretation because a deictic interpretation of the pronoun is more salient in the discourse, then the test can provide little information about knowledge of binding constraints.

It is probably fruitless to try to create scenarios in which the anaphoric and deictic interpretations are perfectly matched in salience. Nevertheless, there are some specific steps that may be taken to minimize the risk of violating Assumption 1.

Guideline 1A. The events in the story that make the deictic interpretation relevant should not overwhelm the events that make the anaphoric interpretation true.

Guideline 1B. Ideally, independent evidence should be provided to show that the scenarios make the anaphoric interpretation accessible, once the possible effects of Principle B are neutralized. This is particularly relevant for tests of bound variable anaphora. If children are reluctant to allow bound variable interpretations for pronouns in general, as has sometimes been suggested (Roeppe 1985, Koster 1994), then children may avoid the anaphoric reading of (18) due to this general bias, independent of Principle B.

Guideline 1C. In a test of the QA, the events that make the anaphoric interpretation accessible (and true) should be maximally similar in the referential and quantificational conditions. If the conditions are poorly matched in this regard, then a spurious QA may be observed.

Elbourne’s (2005) critique of prior studies of the DPBE focuses on Assumption 1. Elbourne suggests that a number of factors independent of Principle B may have led children to avoid the bound pronoun interpretation of sentences like (17) and (18), and that differential presence of these confounding factors may have led to the appearance of a QA. Experiments 1 and 2 below address these concerns by closely matching the test conditions for referential and quantificational conditions, and by independently verifying the availability of bound variable interpretations in children.

Assumption 2: Deictic Interpretation. The scenario should make the deictic interpretation accessible, and should provide a clear motivation for its denial, if it turns out to be false. If a child disallows anaphoric interpretations of (17) and (18) due to Principle B, then the sentences can only be understood as containing deictic pronouns. Therefore, there must be an accessible deictic antecedent for the pronoun him, and the child must be able to judge that the sentence is not true under that interpretation of the pronoun. Importantly, it is difficult to judge that an event did not occur unless it was considered as a live possibility at some point. If an interpretation has never been considered, then children may have difficulty rejecting that interpretation. Furthermore, when a child denies the truth of sentences like (1) and (2), the child’s answer should be consistent with a non-anaphoric interpretation of the pronoun.

Guideline 2A. The deictic interpretation of the test sentences should be accessible. Concretely, this means that the scenario should show that the deictic interpretation was a genuine potential outcome at some point in the story, i.e., it almost became true, and should make it clear why it did not become true. Crain and colleagues have emphasized the importance of this requirement, which they attribute to Russell (1948), and they label it the Condition of Plausible Dissent in TVJT designs. Previous studies of Principle B in children have satisfied – or failed to
satisfy – this requirement in a number of different ways, and we argue below that this factor is important in understanding variability in past findings.

Guideline 2B. In a fair test of the QA, the child’s basis for rejecting the test sentences should be maximally similar in the referential and quantificational conditions. If the same event makes the test sentence false in the two conditions, then the risk of a spurious QA is reduced.

Guideline 2C. Children’s yes/no answers should be followed with requests for explanation. If a child answers ‘no’, but his explanation of his answer indicates that he interpreted the pronoun anaphorically, then this answer clearly provides no evidence for avoidance of anaphoric interpretations.

Assumption 2 is not a focus of Elbourne’s critique, but we suggest that inadequate satisfaction of this assumption may be at least as important as failures to satisfy Assumption 1 in tests of Principle B and the QA in children. Assumption 2 is particularly relevant to the issue of how strong the DPBE is.

3.2 An Example

We next consider in more detail whether the assumptions outlined above are satisfied in one sample TVJT scenario that has been used to motivate both DPBE and the QA. The scenario is drawn from Thornton & Wexler (1999). We focus on this example not because it is better or worse than others, but because the authors provide a detailed description of the study and because it is representative of a design strategy that has been followed in a number of other studies of DPBE and the QA, as discussed further in Section 5 below.

The story in (19) gives the outline of one scenario that Thornton and Wexler used to test sentences with referential and quantificational antecedents alike, following a standard TVJT procedure. The child’s task was to judge Kermit’s statements in (20) or (21).

(19) Bert and three reindeer friends have a snowball fight, and they all get covered in snow. When they go inside, Bert is shivering, so he asks the reindeer to brush the snow off him. Two of the reindeer (separately) refuse, saying they have too much snow to deal with, and they brush themselves. The third reindeer helps Bert a little bit, but then brushes the snow off himself. Bert thanks the helpful reindeer for starting to brush him. He says he’s sorry he can’t reciprocate by helping brush the reindeer; he needs to finish brushing all the snow off himself because he’s still very cold.

(20) I think Bert brushed him. Referential condition
(21) I think every reindeer brushed him. Quantificational condition

The sentences in (20) and (21) are both true in the scenario in (19) only if the pronoun him is interpreted anaphorically, in violation of Principle B. It is true that Bert brushed himself, and also true that every reindeer brushed himself, but neither of these interpretations is acceptable for adults. Thornton and Wexler report that children accepted (20) as true in 58% of trials and accepted (21) as true in only 8% of trials. These results suggest a strong DPBE for referential antecedents and a clear QA.
However, there are a number of potential concerns about how the assumptions behind the TVJT are satisfied here, with specific concerns about whether they are equivalently satisfied in the referential and quantificational conditions.

Consider the first assumption, namely that the anaphoric interpretation is accessible in the scenario, and that only Principle B could prevent a child from accessing this interpretation for the test sentences. Elbourne’s critique focuses on how well this assumption is satisfied, and he suggests that the QA may have arisen simply as a result of identifying the pronoun with the leading character in the story. The story has Bert as the main protagonist, and features three relatively undifferentiated reindeer in largely supporting roles. The main theme of the story is Bert’s search for help. In a scenario of this kind a child may choose Bert as the referent of the pronoun him simply because Bert is by far the most salient male individual in the story. Consequently, a child who has no knowledge of Principle B could reject (21) because the most salient interpretation of the pronoun makes the sentence false (Every reindeer did not brush Bert), and could accept (20) because the most salient interpretation of the pronoun makes the sentence true (Bert did brush Bert). Thus, a QA could have arisen independent of any effect of Principle B. This problem arises because the availability of the anaphoric interpretation is not made parallel across conditions. Elbourne also raises the concern that children may have avoided the bound variable interpretation of (21) due to a general dispreference for bound variable interpretations of pronouns, and not due to Principle B.4

Elbourne’s concerns focus on reasons why children may have chosen the deictic over the bound variable interpretation of the pronoun in (21), independent of Principle B. His arguments question whether any of the guidelines that we listed under Assumption 1 were satisfied. He therefore predicted that appropriately revised experimental tests would show that children would accept local binding with referential and quantificational antecedents alike. Thus, Elbourne’s critique challenges the QA, but assumes that DPBE is a robust finding.

However, if we consider the second assumption behind the TVJT, which focuses on the child’s ability to access and recognize the falsity of a deictic interpretation, then we also find that it is not satisfied as well as might be hoped, leading to concerns about the basic DPBE effect. For the referential condition in (20) the story in (19) clearly motivates an interpretation of the pronoun where him = Bert, i.e., the anaphoric interpretation, but the story does not provide a readily accessible deictic interpretation of the pronoun. Although Bert apologizes that he cannot reciprocate the third reindeer’s generosity by brushing him, this appears not to be an action that was a likely outcome at any point in the story, and it is not clear that the third reindeer is a sufficiently salient referent for the pronoun. As a result, there may have been no suitable deictic interpretation of (20) for a child to deny, violating Guideline 2A and leading to a yes-bias in the referential condition.

Turning to the quantificational condition in (21), the story certainly provides a salient deictic referent for the pronoun. Nevertheless, we might worry that it is never a live possibility that Every reindeer brushed Bert might become true in the story. The first two reindeer reject Bert’s pleas for help, and by the time Bert reaches the final reindeer the suspense in the story surrounds whether even one reindeer will help Bert. Additionally, the events that make the deictic

---

4 Thornton and Wexler show that children willingly accept a bound variable interpretation in a sentence like Every reindeer brushed himself, which replaces the pronoun in (21) with a reflexive. However, the bound variable interpretation is obligatory here, and therefore this does not speak to Elbourne’s concern. Koster (1994) also raises a concern about children’s willingness to accept bound variable interpretations, although the results of Experiment 2 below suggest that this concern is unwarranted.
interpretation of the sentence false differ between the referential and quantificational conditions. In the quantificational condition, the deictic interpretation is made false by the first two reindeer failing to brush Bert; In the referential condition, it is made false by Bert failing to brush one relatively unindividuated reindeer. This asymmetry violates Guideline 2B and contributes further to the concern that the QA might be an experimental artifact.

The concerns about satisfaction of the second assumption appear more serious for the condition with a referential subject NP, thereby raising the possibility that the test may have exaggerated children’s willingness to give non-adultlike judgments for sentences like (20). If this concern generalizes to other studies, and the meta-analysis in Section 5 suggests that in many cases it does, then this provides reason to reexamine children’s knowledge of Principle B, using amended experimental designs and additional controls. In Section 4 we test the consequences of satisfying or failing to satisfy the assumptions that we have discussed in this section.

4. Experiments

In response to Elbourne’s challenge, and in order to address the additional concerns described in Section 3, we conducted three experiments on children’s knowledge of locality constraints on pronominal anaphora. The aim of the experiments was to test whether the Delay of Principle B Effect and the Quantificational Asymmetry persist once the assumptions about TVJT logic outlined above are satisfied. The aim was also to gain better understanding of the substantial variation found in the results of previous studies of QA and DPBE.

Experiment 1 investigates DPBE and QA while providing maximally parallel tests for sentences with referential and quantificational antecedents. Experiment 2 provides an independent measure of the availability of bound pronoun interpretations, by pairing the same scenarios used in Experiment 1 with sentences that are not subject to Principle B. Experiment 3 examines the impact of modifying the scenarios from Experiments 1 and 2 so as to reintroduce some of the concerns raised about previous studies of DPBE and QA. The scenarios used in the experiments are schematized in Figure 1.

4.1 Experiment 1

4.1.1 Design and Participants

Experiment 1 investigated both the DPBE and the QA, using sentences with referential and quantificational subject NPs as in (1) and (2), repeated here as (22) and (23).

(22) Grumpy painted him. Referential condition
(23) Every dwarf painted him. Quantificational condition

The experiment used a Truth Value Judgment Task, in which a child and a puppet companion, Kermit the Frog, watched the experimenter act out a story with props. After the conclusion of the story Kermit made a statement about the story, and the child’s task was to reward or correct Kermit based on the accuracy of his statement.

The experimental materials consisted of 8 stories, each of which was compatible with test sentences from both the referential condition and the quantificational condition. The eight stories were assigned to two lists of items in a Latin Square design, such that each participant saw all 8
stories, paired with 4 referential and 4 quantificational test sentences, and such that across participants each story was paired equally frequently with referential and quantificational test sentences. Participants were randomly assigned to one of the two lists. The 8 target items were combined with 8 filler stories that were intermixed with the target stories to create a test consisting of 16 stories. Filler stories were included to ensure that children would give varied responses, and to provide an independent measure of the children’s understanding of the task. For individual children the stories were divided over 2 sessions of no more than 20 minutes containing 8 stories. Adults were tested in a single session. Participants were 16 English-speaking children aged 4;0-5;6 years (mean age 4;6 years) and 16 adult controls. The age range for the child participants corresponds to the age range that has been claimed to show strong DPBE and QA effects in previous studies. The children were recruited from preschools at the University of Maryland and in the College Park, MD area. Three additional children were replaced in the design because they made errors on more than two filler trials.

We first summarize the plot of a sample story in (24) and then review how the story aims to satisfy the assumptions of the TVJT. All other stories were designed following the same template of events. A full list of test sentences can be found in Appendix A.

(24)  The Painting Story

Characters:    Hiking Smurf, Tennis Smurf, Papa Smurf [collectively Smurfs]  
               Grumpy, Dopey, Happy [collectively dwarves]  

Papa Smurf announces that Snow White is going to have a party, and that she is going to have a painting contest. Papa Smurf declares that he is going to be the judge. Each of the dwarves shows and discusses the color of paint that he is going to use to get painted, as does Tennis Smurf. However, Hiking Smurf does not have any paint, and he wonders whether one of the other characters will be willing to share. He first approaches Happy, who says that he would be glad to help out if any paint remains after he is painted. Fortunately, when Happy is finished some paint remains, and so he paints Hiking Smurf. Hiking Smurf, however, is not yet satisfied, so he approaches Dopey with a similar request, which is similarly successful. Then, Grumpy, who is in such a bad mood that he does not even want to go to the party, declares that he doesn’t need to get painted. The other dwarves really want him to go, and Grumpy agrees to get painted, using all of his paint in the process. After Grumpy is painted, Hiking Smurf approaches him and asks for some paint. Grumpy politely apologizes that he would like to help but cannot, because he has used up all of his paint. Hiking Smurf realizes that his best remaining chance is to ask Tennis Smurf for some extra paint, and Tennis Smurf obliges when he is asked. Finally, everybody is ready for Snow White’s party.

Referential Condition
Kermit: OK, this was a story about painting. Hiking Smurf didn’t have any paint, and Grumpy almost didn’t go to the party. Let me see … I think Grumpy painted him.

5 Although the two groups of characters in each story had a collective name, they were not described as a group using the collective term in Experiments 1 and 2, in order to ensure that they were adequately individuated. In Experiment 3 the collective names were used.
Quantificational Condition

Kermit: OK, this was a story about painting. Hiking Smurf didn’t have any paint, and all the dwarves looked great. Let me see … I think every dwarf painted him.

The story in (24) attempts to satisfy the assumptions underlying the TVJT logic in a maximally similar fashion in the referential and quantificational conditions, as follows.

The same stories were used to test the referential and quantificational conditions, and the within-subjects Latin Square design ensured that differences between the stories themselves could not be responsible for differences in responses in the two conditions. More importantly, the stories were designed such that the same events were the critical determinants of the truth or falsity of the test sentences in the two conditions, reducing the possibility that the salience of any individual character or event might lead to a spurious QA.

As described above, the first key assumption of a TVJT test of Principle B is that the event that supports the anaphoric interpretation of the pronoun is accessible in the story, and that there is no reason for a child to avoid the anaphoric interpretation of the test sentence, except for Principle B. In the story in (24) the most relevant referent for the deictic use of the pronoun is Hiking Smurf, the character who received help from various other characters, but crucially not from Grumpy. Meanwhile, the antecedents for the anaphoric interpretation of the pronoun are Grumpy (referential condition) or the dwarves (quantificational condition). If the story were so strongly focused on Hiking Smurf that he was the only plausible antecedent for the pronoun, then children might avoid the anaphoric interpretation of the pronoun for reasons independent of Principle B (cf. Guideline 1A).

We took three steps to reduce the focus on the deictic antecedent and increase the accessibility of the anaphoric reading, although no explicit reflexives were used in the telling of the story. First, all characters in the story had a clear individual identity in addition to being part of a group. This contrasts with the relatively undifferentiated reindeer in the story in (19). Second, each character draws attention to his need to paint himself, before offering assistance to Hiking Smurf. This increases the availability of the bound variable reading for all of the dwarves, and represents an improvement over (19), in which the bound variable reading was not particularly salient in the quantificational condition. Third, the accessibility of the anaphoric reading in both conditions is increased by the drama surrounding Grumpy’s reluctance to paint himself, which temporarily raises the possibility that both Grumpy painted himself and Every dwarf painted himself might not become true. This is important, since a fair test of the QA requires that the accessibility of the anaphoric reading be matched in the referential and quantificational conditions (Guideline 1C). The event in which Grumpy paints himself ensures that the anaphoric interpretation is made true by exactly the same event in the two conditions. Additionally, both Hiking Smurf and Grumpy/the dwarves are mentioned in the lead-in sentence that precedes the test sentence, although this likely has less impact upon the interpretation of the pronoun than the structure of the story.

Despite our efforts to ensure that nothing other than Principle B might make the anaphoric interpretation of the pronoun inaccessible to the children, it is difficult to prove this using test sentences that violate Principle B. Experiment 2 provides a different test of the accessibility of the bound interpretation, satisfying Guideline 1B.
Figure 1: Schematic illustration of characters and scenarios used in Experiments 1-3, highlighting the events that made the anaphoric interpretation of the pronoun true (left column) and the deictic interpretation of the pronoun false (right column), and comparing the referential condition (dotted blue box) and the quantificational condition (solid red box). Reflexive actions are indicated using curved arrows, transitive actions using straight arrows. An arrow with a cross indicates cases where a character refused to carry out an action. (a) and (b) show that in Experiments 1 and 2 the critical events were maximally similar in the referential and quantificational conditions. In contrast, Experiment 3 modified the scenarios to make them more similar to those used in some previous tests of the QA, where the critical events were not matched in the referential and quantificational conditions, as shown in (c) and (d).

The second key assumption of a TVJT test of Principle B is that children’s negative responses reflect consideration of a deictic interpretation of the pronoun. For the story in (24) this means that children’s negative responses should be based on considering whether the proposition that \{Grumpy/Every dwarf\} painted Hiking Smurf might be true. Both of these eventualities almost becomes true in the story, and a clear reason is provided for why it does not become true. Grumpy’s failure to paint Hiking Smurf is the event that satisfies the Condition of Plausible Dissent (Guideline 2A). More importantly, the same event satisfies this condition in
both the referential and quantificational conditions (Guideline 2B), unlike in (19). Additionally, Grumpy is the last dwarf to consider painting Hiking Smurf. In light of the help offered by the first two dwarves, there is a reasonable expectation that Grumpy should also help Hiking Smurf, and thus it is more clearly almost true that every dwarf painted Hiking Smurf. This contrasts with (19), in which the possibility that every reindeer brushes Bert is taken out of consideration as soon as the first reindeer fails to brush him. These design features were included in order to increase the accessibility of the deictic interpretation of the pronoun, but they provide no guarantee that this is the source of children’s ‘no’ answers. Therefore, we had Kermit probe the child for further explanation after each test sentence, irrespective of the child’s initial response (Guideline 2C). If the child’s explanation of a ‘no’ answer suggested that he had interpreted the pronoun anaphorically, then this response was not counted as evidence for knowledge of Principle B.

Figure 1 provides a schematic illustration of the scenarios used in each of our experiments, highlighting the events that make the anaphoric interpretation of the pronoun true and the events that make the deictic interpretation false, in the referential and quantificational conditions alike.

4.1.2 Results

Results are based on the number of trials on which responses reflected an anaphoric interpretation of the pronoun, which was always true in the story, and the number of trials where responses reflected a deictic interpretation of the pronoun, which was always false in the story. The primary indicator of the pronoun interpretation came from the ‘yes’ and ‘no’ judgments of the puppet’s statements, but if a child’s subsequent justification of his response indicated a different interpretation of the pronoun, this evidence was used to classify the response. For example, if a child gave a ‘yes’ response to the quantificational test sentence every dwarf painted him, but subsequently explained that ‘only those two did’, pointing to the two dwarves who had painted Hiking Smurf, this indicated that the child had interpreted the pronoun deictically. Children’s explanations contradicted their yes/no responses in only 6% of trials (8/128, 6/64 in the quantificational condition).

Results showed that the children and the adult controls consistently avoided the bound interpretation of the pronoun in both conditions. Children accepted the anaphoric interpretation in 11% (7/64) of referential trials and in 14% (9/64) of quantificational trials. This difference was not significant (Wilcoxon Signed Ranks, Z = -0.541; p = 0.59). No child gave more than 2 non-adultlike responses in either condition, and the non-adultlike responses were contributed by 5 children in the referential condition and by 7 children in the quantificational condition. Adult controls accepted the anaphoric interpretation in only 5% (3/64) of referential trials and 3% (2/64) of quantificational trials, again showing no significant difference between conditions (Z = -0.447; p = 0.65). Overall, children accepted more anaphoric interpretations than adults (Wilcoxon Signed Ranks, Z = -2.145, p < .05), but this difference did not interact with experimental condition (Kruskal-Wallis \( \chi^2 = 5.12, p > 0.15 \)).

4.1.3 Discussion

The results of Experiment 1 indicate that children and adults consistently avoided the illicit anaphoric interpretation of the pronoun, instead choosing a deictic interpretation of the pronoun that makes the test sentence false. The choice of deictic interpretation was consistent across
conditions. Assuming that the logic of the TVJT is adequately satisfied, this result suggests that the children and adults avoid the anaphoric interpretation of the pronoun because they respect Principle B. In other words, we find no Delay of Principle B Effect and no Quantificational Asymmetry, contrary to the results of a number of previous studies. However, this conclusion depends on the assumption that the children’s only reason for avoiding the bound interpretation of the pronoun is Principle B, and it is difficult to confirm this step in the argument using materials that are subject to Principle B. Additionally, the results of this experiment are consistent with the concern raised by Elbourne and others that children might exhibit a general dispreference for bound variable interpretations of pronouns, independent of Principle B. We conducted Experiment 2 in order to provide an independent test of whether the bound interpretation of the pronoun is available, once the effect of Principle B is neutralized.

4.2 Experiment 2

4.2.1 Design and Participants

Experiment 2 was designed as a test of whether the anaphoric reading of the pronoun is grammatically available to children, and whether the meaning that supports this reading is sufficiently prominent in the stories used in Experiment 1. Experiment 2 was identical to Experiment 1 in all respects, except for the object NP in the test sentences used after each story. (25) and (26) have identical truth conditions to the illicit anaphoric interpretation of the sentences in (22) and (23), but embedding of the pronoun as a possessor inside the object NP makes the anaphoric readings fully acceptable. If Principle B was the only reason for rejection of the bound interpretation of the pronoun in Experiment 1 then participants should readily accept the bound interpretation in Experiment 2.

(25) Grumpy painted his costume.  Referential condition
(26) Every dwarf painted his costume.  Quantificational condition

Note that the sentences in (25) and (26) provide a fairer test of the availability of the bound reading than does a control in which the pronoun is replaced with a reflexive (e.g., Every dwarf painted himself, cf. Thornton & Wexler 1999). Since a reflexive is obligatorily bound by a local antecedent, a control condition that uses reflexives cannot provide an independent measure of whether a context is equally supportive of the non-bound interpretations of the pronoun.

Participants were 16 English-speaking children aged 4;0-5;4 years (mean age 4;6 years), and 16 adult controls, none of whom had participated in Experiment 1. One child who gave more than two incorrect responses in filler trials was replaced in the design. The children were recruited from preschools at the University of Maryland and in the College Park, MD area. The procedure was identical to Experiment 1.

4.2.2 Results

As in Experiment 1, we used both yes/no responses and children’s justifications of their answers to classify responses as reflecting bound vs. deictic interpretations of the pronoun. Children’s justifications diverged from the default interpretation of yes/no responses in 5% of trials (6/128 trials, 3/64 in the quantificational condition).
The results showed that children accepted the bound interpretation of the pronoun in 80% (51/64) of referential trials and 73% (47/64) of quantificational trials. This difference was not significant (Wilcoxon Signed Ranks \(Z = -0.836; p = 0.40\)). Adults accepted the bound interpretation of the pronoun in 83% (53/64) of referential trials and 67% (43/64) of quantificational trials. This difference was significant (\(Z = -2.640; p < 0.01\)). A comparison of adult and child responses showed no main effect of participant group and no group \(\times\) antecedent interaction. Unsurprisingly, a comparison of the children’s responses in Experiments 1 and 2 showed a highly reliable difference in rates of acceptance of anaphoric interpretations (Kruskal-Wallis \(\chi^2 = 42.395, p = 0.001\)).

4.2.3 Discussion

The only difference between Experiment 1 and Experiment 2 resided in the test questions. In Experiment 1, Principle B was potentially active, whereas in Experiment 2 it was not. By making the pronoun a possessor in Experiment 2, we neutralized any possible contribution of Principle B. The stories and the potential interpretations of the test sentences were otherwise identical to Experiment 1. In light of the dramatic increase in the acceptance of the anaphoric interpretation of the pronoun in Experiment 2, there is good reason to conclude that Principle B was responsible for avoidance of the bound interpretation in Experiment 1. This result leaves little doubt that the test stories make the anaphoric interpretation readily accessible, and that children have little difficulty in accepting bound variable interpretations of pronouns. Adults showed a small but reliable tendency to accept anaphoric readings more frequently in the referential condition. Although this might indicate that the anaphoric interpretation was more salient in the referential condition, it should be noted that this different is proportionally very small compared to the differences that have been presented as evidence for the QA.

Experiments 1 and 2 together satisfy the requirements of a fair TVJT test of DPBE and the QA, and in doing so also address the concerns raised in Elbourne’s (2005) critique. The results lend support to Elbourne’s prediction that appropriately matched referential and quantificational conditions would show no quantificational asymmetry in children. However, our results do not support Elbourne’s further prediction that the improved tests would show a clear DPBE in quantificational and referential conditions alike. In fact, our results show that 4-year old children very rarely chose pronoun interpretations that violate Principle B.

We are now in a position to consider the theoretical implications of the lack of QA and DPBE. The QA has been taken to provide a dramatic piece of evidence in favor of Reinhart’s (1983) approach to binding theory, which restricts the scope of binding constraints to cases of bound variable anaphora. If children do not show a QA this removes one argument in favor of Reinhart’s approach, but it does not provide evidence against this approach, and the theoretical arguments discussed in Section 2.2 are unaffected. However, our findings also have more positive consequences, since they open up theoretical possibilities that may have appeared to be closed off by the QA or the DPBE, and they even remove a potential problem for Reinhart’s theory. In particular, our findings cast doubt upon the surprising asymmetry in children’s mastery of the Principles B and C that has been seen in previous literature. The developmental advantage for Principle C is unexpected under theoretical accounts that we are aware of. Under
Reinhart’s approach both types of disjoint reference effects fall within the scope of Rule I, and therefore no developmental contrast is expected.\(^6\)

However, our 4-year olds’ apparent lack of a DPBE stands in contrast to many previous reports of DPBE in children of a similar age, and this raises the concern that our children’s success may have been due to a fluke or to a design flaw. We took two steps in order to begin to address this concern. First, we conducted an additional experiment (Experiment 3) that was based on the same stories as Experiments 1 and 2, but that reintroduced some of the design features used in previous TVJT tests of DPBE. Second, we conducted a meta-analysis of previous studies of Principle B in children, in an effort to assess the reliability of previous findings and the methodologies used in those studies.

4.3 Experiment 3

4.3.1 Design and Participants

The goal of this experiment was to further test the possibility that the DPBE and the QA observed in previous TVJT studies could have been due to experimental flaws, and to test whether specific features of our stories, or of our population of children, might have artificially exaggerated children’s mastery of Principle B. In order to do this, we made a number of modifications to the stories used in Experiments 1 and 2, making them more like the sample story from Thornton & Wexler (1999), shown above in (19).

The story in (27) is the modified version of the story in (24), and the test sentences are shown in (28-29). We should emphasize that because there are number of differences between the two versions of the story, this experiment cannot identify the exact cause of any difference that might emerge in children’s responses. Rather, it can serve as an initial test of how much the contextual details in a TVJT experiment influence children’s performance on tests involving anaphoric relations.

\begin{align*}
(27) & \quad \text{This is a story about three dwarves and Hiking Smurf. Hiking Smurf announces a} \\
& \quad \text{party at Snow White's house, and declares that everybody needs to get painted for} \\
& \quad \text{the party. He then realizes that he is out of paint, and proceeds to solicit help from} \\
& \quad \text{the dwarves. Hiking Smurf asks the first dwarf to paint him, but he refuses} \\
& \quad \text{because he is too busy painting himself. Hiking Smurf then approaches the second} \\
& \quad \text{dwarf, but he also refuses and paints himself. Hiking Smurf finally asks the third} \\
& \quad \text{dwarf, who is more forthcoming. He says, “I can give you a little of my paint, but} \\
& \quad \text{not too much, I need to get painted”. Hiking Smurf thanks the dwarf and remarks} \\
& \quad \text{that he wishes he could return the favor by helping to paint the dwarf, but cannot} \\
& \quad \text{because he is too busy getting painted himself.}
\end{align*}

\begin{align*}
(28) & \quad \text{Hiking Smurf painted him.} & \text{Referential condition} \\
(29) & \quad \text{Every dwarf painted him.} & \text{Quantificational condition}
\end{align*}

\(^6\) Grodzinsky and Reinhart (1993: pp. 91-93) recognize this discrepancy, and suggest that the evidence for children’s mastery of Principle C is unclear, since studies have often confounded the effects of Principle C with the possible effects of a dispreference for backwards anaphora. We revisit this issue in Section 5, and show that more recent studies have repeatedly confirmed children’s early mastery of Principle C.
In describing how the story in (27) differs from its counterpart in Experiments 1 and 2 we focus on how the core assumptions behind a TVJT test of Principle B and the Quantificational Asymmetry are satisfied. The first assumption is that the story makes the anaphoric interpretation of the pronoun accessible, and that only Principle B could cause the child to avoid this interpretation of the test sentence. The second assumption is that the story places the deictic interpretation under consideration, and makes it clear why this interpretation does not become true. Of course, it is preferable if these desiderata are met in the same way in the referential and quantificational conditions.

The anaphoric interpretations of the test sentences are certainly true in the story in (27): Hiking Smurf eventually paints himself, and the dwarves all paint themselves. However, the focus of the story is not on reflexive actions but rather on Hiking Smurf’s search for somebody who is willing to paint him. As the central protagonist in the story, Hiking Smurf is a strong candidate for a deictic interpretation of the pronoun *him*. This may give children a good reason to avoid the bound interpretation of the pronoun, independent of Principle B (a violation of Guideline 1A). Moreover, this possible bias for a deictic interpretation of the pronoun may impact the referential and quantificational conditions differently. Substituting *Hiking Smurf* for *him* in (28-29) yields the grammatical deictic interpretation in the quantificational condition, but an ungrammatical anaphoric interpretation in the referential condition. This observation reiterates a concern raised by Elbourne (2005) about the possibility of a spurious QA. Additionally, the event that makes the anaphoric reading true differs between the referential and quantificational conditions (a violation of Guideline 1C).

When we turn to the issue of whether the story makes the deictic interpretation accessible and makes it clear why that interpretation does not become true, we find a further contrast between the referential and quantificational conditions. In the referential condition the most relevant deictic interpretation is *Hiking Smurf painted the third dwarf*. Although it is clear that this proposition is false, the third dwarf may be insufficiently prominent to be a viable deictic antecedent for the pronoun, and the proposition is only fleetingly considered (Hiking Smurf says “I wish I could help you, but I can’t”). There is no prior expectation in the story that Hiking Smurf might have painted that dwarf. In the quantificational condition, on the other hand, the most relevant deictic interpretation is *Every dwarf painted Hiking Smurf*. It is also clear that this proposition is false, and in this case Hiking Smurf is a more prominent and plausible deictic antecedent for a pronoun. This creates an imbalance across conditions in how well Guidelines 2A and 2B are satisfied.

In sum, the story in (27) does not satisfy the assumptions underlying the TVJT as fully as its counterpart in (24), and there are a number of differences in how these assumptions are satisfied in the referential and quantificational conditions, increasing the risk of a spurious QA. It should be noted that the story does conform to the basic TVJT parameters of a test of a grammatical constraint, since the story makes an ungrammatical reading true and a grammatical reading false. But as we have tried to emphasize, a fair test requires more than this.

Apart from the changes in the stories, all other details of the design of the experiment were identical to Experiment 1. The 8 test stories were distributed across two lists in a Latin Square design and combined with 8 filler stories to create a task involving 16 stories, which was administered across two testing sessions. Participants were a further 16 English-speaking children aged 4;1-5;2 years (mean age 4;7). Two children were replaced in the design because they gave more than two incorrect responses in the filler trials. The children were recruited from preschools at the University of Maryland and in the College Park, MD area.
4.3.2 Results and Discussion

As in previous experiments we analyzed whether children’s judgments reflected the anaphoric or the deictic reading of the pronoun, using both their yes/no responses and their explanations for their judgments. In the referential condition in this study there was a particularly high number of trials in which children gave ‘no’ responses but gave a justification that suggested a bound interpretation of the pronoun. For example in 30% (19/64) of trials children responded to *Hiking Smurf painted him* by saying something like ‘No, he [pointing to the third dwarf] painted him [pointing to Hiking Smurf]’. This answer suggests that the child may have understood the pronoun in the test sentence to refer to Hiking Smurf, yielding a violation of Principle B, and further suggests that the child did not consider the test sentence to mean *Hiking Smurf painted the third dwarf*. There were no responses of this type in the quantificational condition.

When we combine all trials on which the child’s judgment or justification reflects an anaphoric interpretation we find a clear contrast between the quantificational and referential conditions. Children showed evidence of evaluating the anaphoric interpretation in 56% (36/64) of referential trials, but in only 16% (10/64) of quantificational trials. This difference was highly reliable (Wilcoxon Signed Ranks Z = -2.507; p = 0.01). Table 1 presents a comparison of the rates of acceptance of anaphoric interpretations across all three experiments, showing that only in Experiment 3 was there evidence of a DPBE and a QA.

<table>
<thead>
<tr>
<th>% Accept Binding/Coreference</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Children</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referential Antecedent</td>
<td>7/64 11%</td>
<td>51/64 80%</td>
<td>36/64 56%</td>
</tr>
<tr>
<td>Quantificational Antecedent</td>
<td>9/64 14%</td>
<td>47/64 73%</td>
<td>10/64 16%</td>
</tr>
<tr>
<td><strong>Adults</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referential Antecedent</td>
<td>3/64 5%</td>
<td>53/64 83%</td>
<td></td>
</tr>
<tr>
<td>Quantificational Antecedent</td>
<td>2/64 3%</td>
<td>43/64 67%</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Acceptance rates for anaphoric interpretations of pronouns in Experiments 1-3

The fact that the DPBE and QA emerged with the modified stories used in Experiment 3 lends support to the notion that these effects might have appeared in previous studies due to artifacts of the TVJT designs used. It is impossible to determine the specific cause of the different results in Experiment 1 and Experiment 3, since a number of changes were made to the stories. However, a second asymmetry in the results of Experiment 3 suggests that the children’s worse performance may have been due to insufficient accessibility of the relevant deictic interpretation of the test sentence in the referential condition. We counted the number of trials on which children justified their answers by referring to the rationale for a negative answer that was provided in the story. In the quantificational condition children gave the intended rationale on 32 trials, pointing out that two dwarves did not paint Hiking Smurf or that only the last dwarf did paint Hiking Smurf. In the referential condition, in contrast, there was only 1 trial on which a child gave the intended rationale, e.g., by pointing out that Hiking Smurf was too busy to help the third dwarf. This difference likely reflects the contrasting prominence of the events that satisfy the Condition of Plausible Dissent (Guideline 2A) in the two conditions. There was further evidence that suggested that the children struggled to understand the test sentences in this
experiment. In some trials children responded by simply retelling the story, as if they were not sure of which events were relevant. This occurred on 6 trials in Experiment 3 (9%, 5 referential trials, 1 quantificational trial), and never occurred in Experiments 1-2.

The fact that we were able to recreate the DPBE and QA in Experiment 3 by manipulating the experimental design does not, of course, entail that previous reports of DPBE and QA are also problematic, and there are many such reports. In order to assess the robustness of these findings, we conducted an analysis of previous studies on this topic, described in Section 5.

5. Previous Findings

The classic description of the Quantificational Asymmetry is that children adhere to Principle B with quantificational antecedents, but fail with referential antecedents. This asymmetry mirrors the distinction between two types of anaphoric relations proposed by Reinhart (1983): pronouns can only be linked to quantificational antecedents through bound variable anaphora, whereas pronouns may be related to referential antecedents through accidental coreference. The results of our experiments in Section 4 provide no evidence for a QA and also suggest that the 4-year olds adhere to Principle B. These findings appear to be at odds with a sizeable previous literature on children’s knowledge of Principle B, and for this reason we conducted a meta-analysis of 31 previous studies in light of the various methodological issues that we have discussed here.

We first review the studies that have addressed the existence of a QA by comparing children’s interpretation of pronouns with quantificational and referential antecedents. We find that there is a wide divergence of findings across studies, and that very little evidence for a QA remains once considerations of matching of events and antecedents are taken into account. We then turn to examine the evidence for the DPBE in English and other languages. Here we find that although some studies are subject to methodological concerns, there is good evidence that children do accept interpretations that violate Principle B, albeit at somewhat lower rates than is commonly supposed. In Section 5.3 we compare previous findings on Principles B and C in children and suggest that the ‘residue’ of the DPBE parallels an independently motivated contrast between Principles B and C in how adults access potential antecedents during real-time language processing. In Section 5.4 we consider why children appear to show more adult-like performance in studies of clitic pronouns.

We should note that the variability in findings about the QA and the DPBE has been described in earlier reviews (e.g., Kaufman 1994, Koster 1994, Elbourne 2005). The current meta-analysis is broader and more up-to-date, and it analyzes previous findings in terms of the specific criteria discussed in Sections 3-4. In particular, we focus on the extent to which tests of binding constraints have succeeded in making the interpretations corresponding to the anaphoric and non-anaphoric readings of test sentences readily accessible, such that extra-grammatical accounts of children’s successes and failures can be ruled out.

We begin with a discussion of previous findings concerning the QA.

5.1 The Quantificational Asymmetry

In this domain we examined 18 studies, of which 10 report a quantificational asymmetry and 8 do not.⁷ Across the studies that report a QA and those that do not we find substantial variation

---

⁷ Our meta-analysis includes studies on groups of typically developing monolingual children for which we were able to find at least some details of the methods used. We excluded case studies based on very small numbers of children,
in children’s judgments. The group of studies that fails to find a QA includes a TVJT study by Kaufman (1988) and an act-out study by Lombardi & Sarma (1989), the former reporting that children perform well with both types of antecedents, and the latter reporting that children show equally poor performance with either type of antecedent. Among the studies that do find a QA, the findings are also not consistent across studies. For example, some studies report a QA only within a limited age range (Chien and Wexler 1990, Philip and Coopmans 1996), while others show no age effects (Matsuoka 1997, Savarese 1999). Across the studies on Principle B in children there is broad variation in acceptance rates for illicit referential antecedents, ranging from 10% (Kaufman 1988, 5-year olds) to 82% acceptance (McKee 1985) in studies of English. Therefore, it is important to ask to what extent these diverse findings reflect variation in children’s grammars or variation in the experimental tasks. In what follows, we pay special attention to how closely the referential and quantificational conditions were matched in the different experiments. As we have seen in Sections 3 and 4 it is important in a TVJT test of Principle B to ensure that the event that makes the anaphoric interpretation of the pronoun true and the event that makes the deictic interpretation of the pronoun false are both easily accessible in the test scenario. Further, it is important in a test of the QA to ensure that these events are maximally parallel in referential and quantificational conditions. For example, if the event that makes the deictic interpretation of the pronoun false (thereby satisfying the Condition of Plausible Dissent) is more prominent in the quantificational condition than in the referential condition, then a QA can arise for extra-grammatical reasons. Experiment 3 showed that by introducing an imbalance of this kind we were able to recreate the QA, suggesting that children’s interpretations are highly sensitive to such variables.

We first consider the previous studies that tested for a quantificational asymmetry but did not find one. Some of these studies found that children reliably performed well in referential and quantificational conditions (Kaufman 1988, Hestvik & Philip 1999), whereas some others found that children performed equally poorly with both types of antecedents (Lombardi & Sarma 1989, Boster 1991 Exp. 2, Avrutin & Wexler 1992).

Kaufman (1988) is one of the earliest tests of the QA, and is also possibly the best example of a study that satisfies our criteria for a fair test of the QA. Kaufman used a TVJT in which the scenarios used for the quantificational and referential conditions were very similar in structure. In particular, the deictic antecedent for the pronoun was closely matched in the two conditions, and the event that made the deictic interpretation false was similar across conditions. Kaufman reports almost identical rates of acceptance of around 16% for the two conditions. These results provide evidence against both the QA and the DPBE. In a picture verification task with Norwegian 4-5 year olds Hestvik & Philip (1999) found high rates of success in referential and quantificational conditions alike. Hestvik and Philip report a small difference in acceptance of local anaphora in their two conditions (referential – 3%; quantificational – 9%), and suggest that this reflects a QA. However, these are among the lowest error rates observed in any study of Principle B in children, and they therefore imply early mastery of binding constraints. Hestvik and Philip discuss some grammatical properties of Norwegian pronouns that may have caused the children’s unusually good performance, but the specific cause remains uncertain.

and studies of second language learners or atypically developing children. In cases where an investigator presents multiple studies based on very similar tests, we present only a representative example. Our sample of studies that tested clitic pronouns in Romance and other languages is also not comprehensive, since the finding of improved performance with clitic pronouns is not at odds with the findings in our own studies.

23
In contrast, in a TVJT study with Russian children Avrutin & Wexler (1992) found poor performance with both types of antecedent, with acceptance of anaphoric readings in 40-50% of trials. The scenarios used in this study followed a similar structure to those used by Thornton & Wexler (1999) that we have discussed above. The story revolves around a central character who needs help. He seeks help from each of the members of a group in succession, but is refused by all but the last member of the group. After the central character receives help, he remarks that he regrets that he cannot return the favor. As we have discussed in the context of Thornton and Wexler’s stories and our own Experiment 3, this remark is probably insufficient to place the deictic interpretation of the pronoun under consideration, and therefore it is unsurprising that Avrutin and Wexler found that children showed high levels of acceptance of the anaphoric interpretation of the pronoun. Nevertheless, to the extent that the quantificational condition was similar in this study and Thornton & Wexler (1999), it is unexpected that the Russian children should have shown such high rates of acceptance for the anaphoric reading in the quantificational condition.

Lombardi & Sarma (1989) report an act-out study in which they found similarly high rates of acceptance of anaphoric reading in referential and quantificational conditions alike. Elbourne (2005, p. 355) comments approvingly that ‘this experiment avoided the possible biasing effects of both stories and pictures by using neither’, but consideration of the task used suggests that the pragmatics of the situation could easily have guided the children’s interpretations. Children acted out events in response to requests from a monkey puppet, such as ‘Monkey wants Bert to go into the box and get him a toy’. In this scenario the monkey is the licit non-local antecedent for the pronoun and Bert is the illicit local antecedent. A response was classified as non-adultlike if the child had Bert fetch a toy but not give it to the monkey. Notice, however, that the act of toy-fetching that indicates an ungrammatical reading is a sub-part of the sequence of actions that indicates the grammatical reading: Bert must fetch a toy in order to give it to the monkey. Also, if a child did not attend fully to the test sentence and ignored the pronoun, then the child would have given a response that would be coded as an ungrammatical ‘reflexive’ interpretation. Therefore, it is possible that Lombardi and Sarma’s task may have led to an exaggerated appearance of failure with Principle B.

Thus, the previous literature on children’s mastery of Principle B provides a number of cases where children appear to treat referential and quantificational antecedents similarly. We next turn to studies that have reported a QA. Our sample included 10 studies, of which 6 used the TVJT method, and two each used grammaticality judgment and picture verification tasks. We claim that in many of these experiments, the criteria for an experiment that makes antecedents equally available are not met.

One study that reported a QA used story formats that closely parallel the example from Thornton and Wexler, leading to the same methodological concerns discussed in Section 3.2 (Matsuoka 1997). Another pair of studies that report a QA tested 3-6 year olds using a task that is described as a grammaticality judgment task (McDaniel, Cairns, & Hsu 1990, McDaniel & Maxfield, 1992). These tasks were very similar to tasks that have been described in other studies as TVJTs, and it is not clear that the children consistently understood the difference between a truth-value judgment and an acceptability judgment. Given the similarity with TVJTs, it is very relevant that these tasks did not make the deictic reading of the target sentences accessible. We

8 Lombardi and Sarma’s study tested the four verbs get, give, find, and feed. This concern does not apply to the trials with feed, and therefore cannot straightforwardly account for the 40% ungrammatical interpretations reported for this verb (5/15 referential trials, 7/15 quantificational trials).
discuss these further in a discussion of the importance of the Condition of Plausible Dissent in Section 5.3.

Some other TVJT experiments, such as Thornton (1990), Boster (1991, Exp. 1) and Savarese (1999) appear at first to meet the criteria for a fair test of QA. These studies do more to make both the deictic and anaphoric antecedents accessible in the test stories, and are therefore more likely to satisfy the Condition of Plausible Dissent (but see Section 5.3). However, closer inspection reveals concerns about the evidence that these studies present for a QA.

Thornton (1990) and Boster (1991, Exp. 1) compare children’s judgments of sentences with a referential NP or a \textit{wh}-phrase as a subject, as illustrated in (30-31), and use this comparison as a test of the QA.\footnote{Thornton (1990) uses the examples in (30-31) and the following story to illustrate the design of her study, but these were not, in fact, included in her test of Principle B. The actual stories used are not provided, but they are described as parallel in structure to the example shown here. Both Thornton and Boster included at least some trials with singular pronouns among the test sentences in their studies, e.g., \textit{I know who dressed her} – Baby Sally (Boster 1991).}

\begin{align}
(30) & \quad \text{Bert and Huckleberry Hound scratched them.} \\
(31) & \quad \text{I know who scratched them – Bert and Huckleberry Hound.}
\end{align}

Both of these target sentences can be tested with an identical story. For example, Big Bird, Snuffalupagus, Bert, and Huckleberry Hound take a walk with Robocop and Batman just before dark. Everyone gets bitten by mosquitoes except for Robocop and Batman, who are spared because of their special suits. Big Bird and Snuffalupagus get badly bitten because they are large, and seek help. Robocop and Batman help them out by scratching them, but Bert and Huckleberry Hound refuse to help because they need to attend to their own bites, and they scratch themselves instead. Therefore, it is true that Bert and Huckleberry Hound scratched themselves (anaphoric reading of the plural pronoun), but false that Bert and Huckleberry Hound scratched Big Bird and Snuffalupagus (deictic reading). The logic of the studies as tests of the QA relies on the assumption that the subject NP in (30) is referential, whereas the \textit{wh}-phrase subject in (31) is quantificational. Both studies reveal a very clear contrast between the referential and \textit{wh}-phrase conditions, and thus appear to present strong support for the QA. However, we have two concerns about these tasks. The presence of \textit{who} does not ensure a bound variable interpretation for the pronoun, and there are extra-grammatical interpretive mechanisms that may lead children to respond differently in the two conditions.

In a sentence with a referential subject like (30) a child who does not know Principle B or Rule I could generate an anaphoric interpretation of the pronoun either via variable binding or via accidental coreference. Thornton and Boster reason that when a pronoun has a quantificational antecedent, only the bound variable interpretation is available. However, the sentence in (31) with a \textit{wh}-phrase subject presents exactly the same interpretive possibilities for the pronoun as the sentence in (30). The \textit{wh}-phrase obligatorily binds the gap in subject position, but the \textit{wh}-phrase is not directly or obligatorily linked to the object pronoun \textit{them}. The object pronoun may be treated as a bound variable, linked to the subject, yielding the interpretation \textit{I know which group of individuals $x_1\ldots x_n$ is such that for all $x_i$, $x_i$ scratched $x_i$}, or the pronoun may be treated as referential, allowing accidental coreference, yielding the interpretation \textit{I know which group of individuals $x_1\ldots x_n$ is such that for all $x_i$, $x_i$ scratched Bert and Huckleberry Hound}. Since the interpretive possibilities for (30-31) do not differ on theoretical grounds, it is therefore odd that children should have judged these sentences so differently in Thornton and
Boster’s studies. However, we suggest that there may be an extra-grammatical interpretive strategy that could account for the observed asymmetry.

When a child is presented with the sentence *I know who scratched them*, the child can reasonably infer that the proposition *x scratched them* is presupposed, and that his interpretive task is merely to supply appropriate interpretations for *x* and for *them*. For a child who does not know Principle B a possible interpretation is that *them* is a bound variable pronoun, and therefore that the speaker of the sentence asserts that he knows *which* *x* is such that *x* scratched *x*. However, this is a less likely interpretation, since the possibility of different characters scratching themselves is not a central theme of the story. Alternatively, the child may decide that speaker of the sentence asserts that he knows *which* *x* is such that *x* scratched *Big Bird and Snuffalupagus*. Big Bird and Snuffalupagus are the central characters in the story, and the focus of the narrative is on the fact that one pair of characters is willing to help them whereas another pair of characters is not. This is therefore a very plausible interpretation of the first part of sentence (31), and therefore it should be easy for the child to reject the sentence upon hearing *Bert and Huckleberry Hound*, since they are clearly not the individuals who scratched Big Bird and Snuffalupagus. Consequently, sentence (31) may not indicate that the child’s grammar excludes the bound variable interpretation of the pronoun, since that interpretation may never have been under consideration. What appeared at first to be an elegant test of the QA may again indicate the need for very careful attention to the pragmatics of the test protocols.  

Savarese (1999) presents another TVJT test of the QA that appears to satisfy the Condition of Plausible Dissent and makes both the deictic and anaphoric antecedents accessible within each condition. The format of his stories is similar to those used in Kaufman (1988). However, since Savarese’s study takes the interesting step of using the negative quantifier *no* in his quantificational condition, it was impossible to closely match the materials for his referential and quantificational conditions, unlike the stories in our experiments, which were identical across conditions. Additionally, Savarese reports a curious additional asymmetry that is not predicted by Binding Theory. He observed 31% acceptance of coreference in his main referential condition, but in a further condition that included both a referential subject NP and negation (e.g., *Mama Bear didn’t dry her*) Savarese found only 9% acceptance of coreference. This raises the possibility that the low acceptance rates observed in his studies are related to the presence of negation, rather than to a QA.  

---

10 Note that these concerns apply specifically to the use of *wh*-extraction as a test of the QA, and they do not automatically generalize to studies where a similar experimental strategy has been used to test other constraints. Thus, the use of *wh*-phrases may still provide ingenious tests of Strong Crossover (Crain & Thornton 1998).

11 One further TVJT study on English by Avrutin and Thornton (1994) presents a QA-like effect as evidence for a theory in which collective and distributive readings of pronouns map onto coreferential vs. bound variable representations, respectively. Avrutin and Thornton report that many children accepted the illicit anaphoric reading of *The smurf and the clown dried them* more frequently in situations where the clown and the smurf dried themselves collectively than in situations where they dried themselves in separate events. In both conditions the clown and the smurf refused, either jointly or on separate occasions, to dry a pair of other characters. However, the contrast between the collective and distributive conditions may be captured without the need to appeal to a QA. First, in the collective condition the children showed a remarkably strong bimodal distribution in responses. Of the 33 children tested in the collective condition, 17 children showed a strong preference for the deictic interpretation of the pronoun, responding ‘no’ on 68/68 trials (100%), and the remaining 16 children showed a strong preference for the illicit anaphoric interpretation, responding ‘yes’ on 60/64 trials (94%). Although the children as a group showed close to 50% acceptance of anaphoric interpretations, the split in the results is uncharacteristic of studies of this kind, and is inconsistent with theories that attribute DPBE to some kind of uncertainty in children’s judgments (e.g., Chien & Wexler 1990, Grodzinsky & Reinhart 1993). Second, when the 16 children who accepted anaphoric
It is important to also comment here on the relation between the TVJT method that we have discussed at length here and the picture selection or picture verification tasks that have been used in many studies of Principle B in children, including the most well-known report of a QA (Chien & Wexler 1990) and a number of other studies of Principle B in children (e.g., Philip & Coopmans 1996, Hamann, Kowalski & Philip 1997). We contend that picture-based tasks are subject to the same constraints that we have discussed for TVJTs, but that it is more difficult in picture-based tasks to assess how well the constraints are satisfied. In a picture verification task, children are placed in a situation where they could choose to interpret a pronoun either deictically or anaphorically. This choice may, of course, be constrained by their grammar, but many other things can affect the choice. This is exactly like the TVJT setting, except that the TVJT involves a question about a narrative, rather than a question about a picture. We have seen in the current experiments and in many other TVJTs used in our work that children are highly sensitive to the structure of a story, to the prominence of its characters, and to expectations about what events are likely to be commented upon. We have also seen that these extra-grammatical factors can modulate children’s choices of pronoun reference. It is a strength of the TVJT that it encourages the experimenter to make very explicit assumptions about the context in which a child judges a test sentence. Picture-based tasks present similar issues, but we know relatively little about how the specifics of a static picture create a context that can guide a child’s interpretation of a pronoun. The same factors that can lead to extra-grammatical biasing of pronoun interpretation in TVJTs can easily do the same in picture tasks, where the experimenter has even less control over how a child might interpret the scenario placed in front of him. In Chien and Wexler’s classic study children were shown a line drawing and told “Here is Goldilocks, and here are the three bears. Is every bear touching her?” How are we to know how a child might interpret the pronoun if Principle B were irrelevant? It is difficult to answer this question, since we do not have a good theory of how children construct contexts from simple drawings. For this reason it is difficult to evaluate the extent to which picture-based studies represent fair tests of children’s knowledge of Principle B. However, we have no reason to believe that the picture-based studies conjure up contexts for the children that are more balanced than those created in TVJT studies. The reader is referred to Elbourne (2005) for more specific discussion of picture-based tests of the QA.

In sum, our meta-analysis of previous studies indicates that the results of our three experiments do not conflict with the previous literature. Many previous studies have also failed to find a QA, and most studies that have reported a QA are amenable to alternative explanations that do not invoke a grammatical asymmetry between coreference and variable binding. Recall that our Experiment 3 showed that it was possible to recreate a QA by modifying our stories to more closely resemble the scenarios used in some previous studies reporting a QA.

We should emphasize that despite our criticisms of particular TVJT studies, we do not take these examples to show that the TVJT is fundamentally flawed, or that other experimental measures are superior. We believe that our critique and our own studies follow closely the underlying logic of the TVJT, as laid out by Crain and his colleagues. We consider it a strength

interpretations in the collective condition were subsequently tested in the distributive condition, the rate of acceptance fell to 42% (29/64 trials). This drop in acceptance rates may reflect a reluctance to interpret the plural pronoun *them* as a bound variable rather than a specific effect of Principle B (but cf. Thornton 1990, p. 165), but it could also reflect story-specific effects. The article implies that children were tested on closely matched stories in the collective and distributive conditions, which would have led to a test-retest confound. A more likely situation, based on other studies by the same authors (e.g., Thornton 1990) is that the children were tested on entirely different stories in the two conditions, raising the possibility of story-specific biases in the results.
of the TVJT that it encourages the experimenter to make very explicit assumptions about the context in which a child judges a test sentence. Our criticisms apply to specific studies, and not to the task itself.

<table>
<thead>
<tr>
<th>Author</th>
<th>Language</th>
<th>Age</th>
<th>N</th>
<th>Accept Referential</th>
<th>Accept Quant.</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaufman 1988</td>
<td>English</td>
<td>2:7-3:11</td>
<td>30</td>
<td>23%</td>
<td>18%</td>
<td>TVJT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5:0-6:5</td>
<td>30</td>
<td>10%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>Lombardi &amp; Sarma</td>
<td>English</td>
<td>4:0-6:2</td>
<td>11</td>
<td>55%</td>
<td>49%</td>
<td>Act out, TVJT</td>
</tr>
<tr>
<td>1989</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boster 1991, Exp. 2</td>
<td>English</td>
<td>3:3-4:9</td>
<td>24</td>
<td>38%</td>
<td>42%</td>
<td>Picture Verification</td>
</tr>
<tr>
<td>Avrutin &amp; Wexler</td>
<td>Russian</td>
<td>4-7</td>
<td>16</td>
<td>52%</td>
<td>41%</td>
<td>TVJT</td>
</tr>
<tr>
<td>1992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utakis 1995</td>
<td>English</td>
<td>3:4-9:5</td>
<td>30</td>
<td>37%</td>
<td>40%</td>
<td>TVJT</td>
</tr>
<tr>
<td>Baauw, Escobar &amp;</td>
<td>Spanish; clitic</td>
<td>mean 5:6</td>
<td>45</td>
<td>10%</td>
<td>10%</td>
<td>Picture</td>
</tr>
<tr>
<td>Philip 1997</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Verification</td>
</tr>
<tr>
<td>Hamann, Kowalski &amp;</td>
<td>French; clitic</td>
<td>3:5-4:8</td>
<td>9</td>
<td>22%</td>
<td>30%</td>
<td>Picture</td>
</tr>
<tr>
<td>Philip 1997</td>
<td></td>
<td>5:3-5:11</td>
<td>8</td>
<td>0%</td>
<td>12%</td>
<td>Verification</td>
</tr>
<tr>
<td>Hestvik &amp; Philip</td>
<td>Norwegian</td>
<td>4:5-5:11</td>
<td>15</td>
<td>9%</td>
<td>3%</td>
<td>Picture</td>
</tr>
<tr>
<td>1999</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Verification</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author</th>
<th>Language</th>
<th>Age</th>
<th>N</th>
<th>Accept Referential</th>
<th>Accept Quant.</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chien &amp; Wexler</td>
<td>English</td>
<td>2:6-3:11</td>
<td>48</td>
<td>70%</td>
<td>54%</td>
<td>Picture</td>
</tr>
<tr>
<td>1990, Exp. 4</td>
<td></td>
<td>4:0-4:11</td>
<td>45</td>
<td>60%</td>
<td>40%</td>
<td>Verification</td>
</tr>
<tr>
<td>McDaniel, Cairns &amp;</td>
<td>English</td>
<td>2:9-6:7</td>
<td>19</td>
<td>44%</td>
<td>19%</td>
<td>Grammaticality</td>
</tr>
<tr>
<td>Hsu 1990</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td>Thornton 1990</td>
<td>English; “who”</td>
<td>3:7-4:8</td>
<td>12</td>
<td>49%</td>
<td>8%</td>
<td>TVJT</td>
</tr>
<tr>
<td>Boster 1991, Exp. 1</td>
<td>English; “who”</td>
<td>4:6-6:0</td>
<td>10</td>
<td>38%</td>
<td>4%</td>
<td>TVJT</td>
</tr>
<tr>
<td>McDaniel &amp; Maxfield</td>
<td>English</td>
<td>3:1-6:10</td>
<td>37</td>
<td>41%</td>
<td>25%</td>
<td>Grammaticality</td>
</tr>
<tr>
<td>1992</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Judgment</td>
</tr>
<tr>
<td>Avrutin &amp; Thornton</td>
<td>English; collective vs. distributive</td>
<td>3:10-4:10</td>
<td>33</td>
<td>93%</td>
<td>27%</td>
<td>TVJT</td>
</tr>
<tr>
<td>1994</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philip &amp; Coopmans</td>
<td>English</td>
<td>3:6-7:0</td>
<td>19</td>
<td>68%</td>
<td>26%</td>
<td>Picture</td>
</tr>
<tr>
<td>1996</td>
<td>Dutch; strong pronoun</td>
<td>4:3-6:11</td>
<td>37</td>
<td>66%</td>
<td>50%</td>
<td>Verification</td>
</tr>
<tr>
<td>Matsuoka 1997</td>
<td>English</td>
<td>3:10-6:0</td>
<td>19</td>
<td>70%</td>
<td>20%</td>
<td>TVJT</td>
</tr>
<tr>
<td>Savarese 1999</td>
<td>English</td>
<td>3:5-5:11</td>
<td>25</td>
<td>31%</td>
<td>N/A</td>
<td>TVJT</td>
</tr>
<tr>
<td>Thornton &amp; Wexler</td>
<td>English</td>
<td>4:0-5:1</td>
<td>19</td>
<td>58%</td>
<td>8%</td>
<td>TVJT</td>
</tr>
</tbody>
</table>

**Table 2:** Summary of results from tests of the Quantificational Asymmetry in children.
5.2 The Delay of Principle B Effect

We now turn to the question of whether there is a reliable Delay of Principle B Effect. Our results, together with those of Kaufman (1988), suggest that there is no DPBE, and that children perform well across all types of antecedents. However, this conclusion is at odds with many previous studies. How can we reconcile our results with previous studies that report a DPBE? In this section we focus on the empirical record, and conclude that the DPBE is a real effect and is not an experimental artifact, but we also show that the strength of the effect varies greatly as a function of experimental design. In studies that satisfy our key criteria for fair tests of DPBE we generally find that anaphoric readings are accepted in 15-30% of trials, which is lower than is commonly assumed. Although this rate of acceptance is around 10-15% higher than is observed in tests of Principle C, it also does not support the idea that children fail outright with respect to Principle B. In Section 5.3 we turn to the question of what is responsible for the residual DPBE effect, and why children perform differently in tests of Principle B and Principle C. We suggest that this small difference mirrors differences found in measures of on-line reference resolution in adults. In Section 5.4 we suggest further that this difference does not arise in children who speak clitic-languages, due to a difference in the syntactic categorization of pronouns.

In addition to the 18 studies that tested the QA, our survey included a further 13 studies that tested children’s adherence to Principle B with referential antecedents only. The results of these studies are summarized in Table 3. If we restrict our attention initially to studies on English, this leaves 12 studies from Table 2 and an additional 7 studies from Table 3 that report a DPBE, with acceptance rates for local antecedents of pronouns that range from 16% to 82%. This high degree of variability in acceptance rates suggests that children’s responses are not simply the product of a relatively stable grammar. If each study presented a test of Principle B that was immune to extra-grammatical biases, then we should have expected to observe more consistent results across studies. This implies that the variability is due to the specifics of the tasks used in individual studies.

We are not able to present here a detailed analysis of all individual studies of the DPBE, both because of space constraints and because many of the previous studies provide insufficient experimental detail to allow evaluation of the studies. Nevertheless, we can offer comments on possible sources of variability observed in some studies, particularly in the 9 TVJT studies that have reported a DPBE, since they can be more directly compared to the new experiments presented above.

Overall, TVJT tests of DPBE can be constructed more easily than tests of the QA, because basic DPBE tests require fewer factors to be matched in the experimental design. Nevertheless, the same assumptions that underlie a test of the QA also underlie a test of the DPBE. That is, it is important to ensure that both deictic and anaphoric interpretations of the pronoun are accessible in the story, even if they do not ultimately turn out to be true. This requires such measures as avoiding making one character the central focus of the story, and ideally providing an independent measure of the availability of these readings (as in our Experiment 2).

In all TVJT tests of Principle B that we are aware of the child is presented with a situation in which an (ungrammatical) anaphoric interpretation of the pronoun is true and an alternative deictic interpretation of the pronoun is false. In order to reject the target sentence as false, a child must be able to avoid the anaphoric interpretation (perhaps due to Principle B) and access the deictic interpretation and recognize that it is false in the story. Just as we found that it is crucial in tests of the QA to match the accessibility of the deictic interpretation across conditions, our
survey of tests of DPBE suggests that the accessibility of the deictic interpretation is an important predictor of how frequently children accept ungrammatical anaphoric interpretations. We find at least four ways in which studies address, or fail to address, this requirement, which is covered by our Guideline 2A, otherwise known as the Condition of Plausible Dissent.

In some TVJT studies of DPBE the event corresponding to the deictic interpretation of the pronoun never comes under consideration in the scenario. These studies tend to report high rates of anaphoric interpretations. One such study is McKee (1982), the study with the highest acceptance rate in our sample (82%). An example of one of McKee’s stories is the following, involving a princess and a cabbage patch baby. The princess falls into a tub and gets wet. The cabbage patch baby says ‘You’re wet’ and then leaves, after which the princess dries herself. After this story, children were asked to judge the target sentence The princess dried her. Although there is a possible deictic interpretation of the pronoun in this sentence that makes the sentence false, the story does little to make that interpretation accessible. The only washing event that is ever considered in the story is the princess washing herself. This likely made it difficult for children to consider a deictic interpretation of the pronoun and reject the sentence. Similar stories in which the deictic interpretation is not considered are used by McDaniel, Cairns, & Hsu (1990; 44% acceptance) and McDaniel & Maxfield (1992; 41% acceptance). These studies are described as Grammaticality Judgment tasks, but they are so similar to TVJT’s that the children may have given truth judgments rather than well-formedness judgments. Another study in which the deictic interpretation is not considered in the scenario is Grimshaw & Rosen (1990; 42% acceptance). Grimshaw and Rosen used a pared-down version of the TVJT in which children judged sentences that were paired with simple movie clips. For example, in one of their two trials that tested Principle B violations, children saw a movie clip in which Big Bird and Ernie stood next to each other and Big Bird hit himself. Children were then told “Big Bird was standing with Ernie. Big Bird hit him.” This mini-discourse is compatible with a deictic interpretation of the pronoun, in which him refers to Ernie, but the possibility that Big Bird might hit Ernie is never entertained in the scenario. This factor may be particularly important for children, leading them to consider the anaphoric interpretation in a substantial proportion of trials. The same factor may exert a less powerful influence on adults, for whom it is natural to interpret the pronoun in the target sentence as referring to Ernie.

In a second set of studies the event corresponding to the deictic interpretation of the pronoun is explicitly avoided, but the story creates no expectation that this event might happen, and the deictic interpretation is at odds with the focus of the story. These studies provide at most weak satisfaction of the Condition of Plausible Dissent, and they yield high rates of acceptance of anaphoric interpretations. A number of studies of the QA fall into this category, and they have already been discussed above (Thornton & Wexler 1999: 58% acceptance; Matsuoka 1997, 70% acceptance; our own Experiment 3, 56% acceptance).

In a third set of studies the Condition of Plausible Dissent is more clearly satisfied. In these studies the deictic interpretation of the pronoun corresponds to the focal character(s) in the story, and the event corresponding to the deictic interpretation is explicitly avoided in the story, as when a ‘selfish’ character refuses to assist the ‘helpless’ character at the center of the narrative. In these studies there is variation in whether the story creates an expectation that the deictic interpretation might become true. These studies also tend to report lower rates of acceptance of

---

12 It is nevertheless very interesting that McKee’s study reports a much lower acceptance rate (15%) in a variant of the same task presented to Italian-speaking children. We return to this issue below.

13 Interestingly, a very similar point is raised by Crain & McKee (1987, cited in Thornton 1990, p. 169).
the anaphoric reading of the pronoun (Kaufman 1988, 16% acceptance\(^{14}\); Thornton 1990, 29% acceptance across conditions; Boster 1991, 21% acceptance across conditions; Savarese 1999, 31% acceptance; Kiguchi & Thornton 2004, 27% acceptance).

Finally, our own Experiment 1 (11% acceptance) provided scenarios in which there is a clear expectation that the event corresponding to the deictic interpretation might become true. After Hiking Smurf receives help from the first two dwarves, there is a reasonable expectation that Grumpy might be similarly helpful. Grumpy’s refusal is an unexpected turn of events, and therefore the Condition of Plausible Dissent is clearly satisfied. (Our stories also make the anaphoric interpretation similarly salient, since Grumpy almost refused to paint himself.)

Thus, the varying results of previous TVJT studies with English-speaking children bear out our claim about the importance of the Condition of Plausible Dissent in tests of Principle B.

Our comments on studies that have used other experimental measures are more brief, since they have already been raised in the discussion of the QA in Section 5.1. Picture verification tasks use a similar logic to TVJT tests of Principle B, with the difference that it is harder to evaluate the context against which the child judges the test sentence. The three picture verification tests on English in our sample show that illicit anaphoric interpretations of pronouns are accepted at high rates, ranging from 38% to 68% (Chien & Wexler 1990 Exp. 4, Boster 1991, Philip & Coopmans 1996).

Wexler & Chien (1985) used a picture selection task, in which children listened to sentences like Cinderella’s sister points to her and had to find a picture that showed the scenario described in the sentence. In this task the children chose pictures that indicated Principle B violations on 43% of trials. This task is different from a TVJT in the respect that the child’s task is to find which picture makes the test sentence true, rather than to determine whether the test sentence is true. This makes such requirements as the Condition of Plausible Dissent irrelevant to this task, but it does not remove the need for a viable antecedent for the pronoun. We do not have access to the materials for this study, but we tentatively suggest that uncertainty over the deictic antecedent of the pronoun may have made the children more susceptible to ungrammatical interpretations. The familiarization phase for each trial presumably did not make either Cinderella or her sister a more likely deictic center, and the focus of the test pictures was on what the sister was doing: pointing to herself in the ‘reflexive’ picture, and pointing to Cinderella in the ‘pronoun’ picture. In this contrastive situation the most natural description of the ‘pronoun’ picture may be to use the direct object Cinderella rather than the pronoun her. However, we do not know how sensitive children might be to this felicity condition.

A number of early studies of DPBE used act-out tasks (Jakubowicz 1984, Solan 1987, Lombardi & Sarma 1989). These tasks have the limitation that they track a child’s preferred interpretation of a test sentence, and cannot readily distinguish dispreferred from illicit interpretations. However, a distinct advantage is that the child’s act-out provides more direct evidence of his interpretation than does the yes/no response used in TVJT and picture verification tasks. Jakubowicz (1984) used an act-out task in one of the earliest demonstrations of the DPBE, and showed a relatively low rate of anaphoric interpretations of the pronoun (25-30%). Chien & Wexler (1990, Exps 1-2) also found a relatively low rate of anaphoric interpretations (29%) in their act-out task conducted with around 300 children. Rather higher rates of anaphoric interpretations have been reported in other act-out studies, but we have seen

\(^{14}\) Kaufman’s descriptions suggest that the events corresponding to the deictic interpretation of the pronoun were explicitly avoided by characters in her stories, but it is not clear whether this was a consistent feature of the stories.
above that in at least one of these cases the design may have led to exaggeration of the number of anaphoric interpretations (Lombardi & Sarma 1989).

Summarizing previous studies on DPBE in English, we find that DPBE is weaker than often supposed. In the studies that we take to present the fairest tests of DPBE we find that children accept illicit anaphoric readings of pronouns in only 15-30% of trials. We do not find evidence that children ‘guess’ when presented with potential Principle B violations, nor do we find evidence that children misanalyze pronouns as elements that require or strongly prefer local binding. In this respect, the results from our own experiments are consistent with previous literature.

<table>
<thead>
<tr>
<th>Study</th>
<th>Language</th>
<th>Age</th>
<th>N</th>
<th>Accept Ref.</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>11</td>
<td>25%†</td>
<td></td>
</tr>
<tr>
<td>Wexler &amp; Chien (1985)</td>
<td>English</td>
<td>2;6-6;6</td>
<td>129</td>
<td>43%</td>
<td>Picture Selection</td>
</tr>
<tr>
<td>Deutsch, Koster &amp; Koster (1986)</td>
<td>Dutch; strong pron.</td>
<td>6</td>
<td>32</td>
<td>46%</td>
<td>Picture Selection</td>
</tr>
<tr>
<td>Solan (1987)</td>
<td>English</td>
<td>4-7</td>
<td>37</td>
<td>57%</td>
<td>Act Out</td>
</tr>
<tr>
<td>Chien &amp; Wexler (1990), Exps 1-2</td>
<td>English</td>
<td>2;6-6;6</td>
<td>298</td>
<td>29%†</td>
<td>Act Out</td>
</tr>
<tr>
<td>Grimshaw &amp; Rosen (1990)</td>
<td>English</td>
<td>4-5</td>
<td>12</td>
<td>42%</td>
<td>TVJT</td>
</tr>
<tr>
<td>Padilla (1990)</td>
<td>Spanish; clitic</td>
<td>3;0-3;11</td>
<td>20</td>
<td>37%</td>
<td>Act out</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5;0-5;11</td>
<td>20</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>McKee (1992)</td>
<td>English</td>
<td>2;6-5;3</td>
<td>60</td>
<td>82%</td>
<td>TVJT</td>
</tr>
<tr>
<td></td>
<td>Italian; clitic</td>
<td>3;7-5;5</td>
<td>30</td>
<td>15%</td>
<td>TVJT</td>
</tr>
<tr>
<td>Sigurjónsdóttir &amp; Hyams (1992)</td>
<td>Icelandic</td>
<td>4;0-4;6</td>
<td>10</td>
<td>45%</td>
<td>TVJT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4;6-5;0</td>
<td>10</td>
<td>43%</td>
<td></td>
</tr>
<tr>
<td>Baauw (1999)</td>
<td>Dutch; weak pron.</td>
<td>4;2-5;3</td>
<td>15</td>
<td>47%</td>
<td>Picture Verification</td>
</tr>
<tr>
<td>Varlokosta (2000)</td>
<td>Greek; clitic</td>
<td>3;7-5;6</td>
<td>20</td>
<td>13%</td>
<td>TVJT</td>
</tr>
<tr>
<td></td>
<td>Greek; strong pronoun</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Greek; clitic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varlokosta &amp; Dullaart (2001)</td>
<td>Greek; strong pron.</td>
<td>3;3-7;6</td>
<td>10</td>
<td>5%</td>
<td>TVJT</td>
</tr>
<tr>
<td></td>
<td>Dutch; weak pron.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dutch; strong pron.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiguchi &amp; Thornton (2004)</td>
<td>English</td>
<td>4;1-5;10</td>
<td>13</td>
<td>27%</td>
<td>TVJT</td>
</tr>
</tbody>
</table>

**Table 3:** Summary of selected tests of the Delay of Principle B Effect, excluding studies already covered in Table 2 by the survey of tests of the Quantificational Asymmetry. † The percentages for Jakubowicz (1984) are estimates derived from published histograms; the average shown for Chien & Wexler (1990) is a non-weighted average derived from the mean of all age groups.

However, we cannot conclude from the meta-analysis of DPBE studies that the effect is artifactual. Even in studies that present fair tests of binding constraints we find that some form of DPBE remains in many studies. We cannot easily write off 20-30% violations as ‘noise’ in the data, particularly since studies of other grammatical constraints in children commonly report
rather lower error levels. Nor can we simply set aside the studies that we criticized on methodological grounds, for it remains interesting that certain types of experimental settings induce children to accept ungrammatical anaphoric interpretations of a pronoun at high rates. If Principle B acts as a strong constraint on children’s interpretations, then we might expect that Principle B should be sufficiently powerful to make children ‘blind’ to illicit anaphoric interpretations of pronouns, but this appears to not be the case. In addition, we must also consider the fact that some studies in other languages have shown that similar or identical tests elicit much lower rates of Principle B violations. In particular, a number of studies on languages with clitic pronouns, such as Italian and Spanish, have found much lower rates of acceptance of anaphoric interpretations than their English counterparts have. In the next sections we consider the source of the residual DPBE effect, and why children behave differently in tests of Principle C and in tests involving clitic pronouns.

5.3 Principle B vs. Principle C

Although our meta-analysis supported the conclusion from our own experiments that the QA is an experimental artifact, our conclusions about DPBE are more nuanced. Our own studies indicated that Principle B had a strong impact on 4-year olds’ judgments, since the children very rarely accepted anaphoric interpretations in Experiment 1, where Principle B was relevant, and very frequently accepted anaphoric interpretations in Experiment 2, where Principle B was neutralized. However, we were still left with a number of studies that appear to present fair tests of children’s knowledge of Principle B, and that showed acceptance of illicit antecedents on ~15%-30% of trials. This is a weaker DPBE than is often assumed in the literature (e.g., Reinhart 2006), but it cannot easily be dismissed. We also cannot ignore the fact that certain experimental settings make children more likely to accept illicit anaphoric interpretations.

Although there is sure to be a certain level of noise in judgments collected from 4-year olds, it is difficult to attribute the non-adultlike responses in Principle B studies to experimental noise, since many TVJT tests of other constraints show lower error rates. A useful comparison can be found in TVJT studies of Principle C, which have typically shown error rates of around 10-20% (Crain & McKee 1985, 12% acceptance; Guasti & Chierchia 1999/2000, 11% acceptance; Kazanina & Phillips 2001, 17% acceptance). These findings have been taken to indicate that children know Principle C by the age of 3-5. See also Lust, Loveland and Kornet (1980), Solan (1983), McDaniel, Cairns and Hsu (1990) for studies using other techniques, and Lust, Eisele and Mazuka (1992) for a review of earlier studies on this topic.

Therefore, the ‘residue’ of the DPBE appears to be slightly elevated rates of accepting illicit antecedents in Principle B contexts, relative to Principle C contexts. We cannot reasonably argue from such small differences that 4-year olds know Principle C but do not know Principle B (or that they do not know Rule I and related constraints). It has sometimes been suggested that children appear to perform better on tests of Principle C because of a general dispreference for

---

15 Grimshaw & Rosen (1990) report a TVJT study that showed a much higher rate of acceptance of anaphoric interpretations. Children watched movie clips and judged statements about the clips. For example, in scene that shows Ernie hitting himself, children were told "Ernie was fighting with Big Bird. He hit Ernie." Grimshaw and Rosen’s description of their study suggests that they used a strict coding scheme in which all ‘yes’ and ‘no’ answers were considered as relevant to the experimental hypothesis, even when the children’s answers suggested otherwise, such as when children gave a negative response, because hitting isn’t right. It is therefore possible that some of the children’s ‘yes’ answers reflected an inference from the lead-in sentence that mentioned that Ernie and Big Bird were fighting, rather than acceptance of interpretations that violate Principle C.
backwards anaphora. However, a number of studies of Principle C have shown that children freely accept backwards anaphora once the effect of Principle C is neutralized (Crain & McKee 1985, Guasti & Chierchia 1999/2000, Kazanina & Phillips 2001), and therefore it is difficult to dismiss the low error rates in tests of Principle C as experimental artifacts. We must therefore look elsewhere for an explanation of children’s slightly degraded performance on tests of Principle B.

One possible explanation derives from a similar contrast between Principle B and Principle C that has been observed in recent on-line studies of pronoun resolution in adults. These studies have asked whether binding constraints act as an ‘initial filter’ on the processing of pronouns, such that the parser is blind to potential antecedents in grammatically illicit positions, or whether they act as a ‘late filter’, such that comprehenders temporarily consider grammatically illicit antecedents for a pronoun. Existing evidence suggests that Principle C acts as an initial filter, such that comprehenders do not attempt to link pronouns to R-expressions that they c-command (Cowart & Cairns 1987, Kazanina et al. in press), whereas the results for Principle B are rather more mixed. Some studies using cross-modal priming and self-paced reading methods have presented evidence that Principle B acts as an initial filter (Nicol & Swimney 1989, Clifton et al. 1997, Lee & Williams 2006), but a number of more recent studies using eye-tracking and self-paced reading measures have found evidence for temporary consideration of ungrammatical antecedents in Principle B contexts (Badecker & Straub 2002, Kennison 2003, Runner et al. 2003ab, 2006, Sturt et al. 2005).

Although the adult results indicate merely fleeting access to ungrammatical antecedents in on-line studies, whereas the results from children indicate ‘off-line’ judgments that violate Principle B, there is good reason to think that these might be related. A recurring finding in studies of children’s language processing is that children show much greater difficulty than adults in inhibiting and recovering from incorrect initial interpretations of sentences (e.g., Hamburger & Crain 1984, Trueswell et al. 1998). Therefore, what appears in adults as transient effects of ungrammatical antecedents might appear in children as ungrammatical interpretations that are longer lasting.

Next, we can ask why Principle B and Principle C should impact the on-line search for pronoun antecedents in different ways. Although we cannot exclude the possibility that the constraints themselves are qualitatively different from one another, there are independent reasons why the search for antecedents might proceed differently in the two cases. These differences are all related to the fact that Principle B primarily constrains forwards anaphora, whereas Principle C primarily constrains backwards anaphora. In backwards anaphora contexts a pronoun precedes its antecedent, and encountering a pronoun initiates an active search for a suitable antecedent (Kazanina et al. in press). During this search, the parser is able to consider potential antecedents one at a time as they appear in the input, with no need to retrieve antecedents from memory. Additionally, the parser can identify that a given structural domain cannot contain an antecedent for the pronoun, due to Principle C, before it encounters any of the NPs in that domain. In contrast, the forwards anaphora contexts that are normally used in tests of Principle B place different demands on the anaphora resolution process. The parser encounters the pronoun only after it has encountered its potential antecedents, and must therefore conduct a retrospective search of referents in memory. Furthermore, the contexts that are typically used in tests of Principle B in children and adults force the parser to consider multiple candidate antecedents (intrasedentential or extrasedentential) in parallel. Both of these factors may increase the likelihood of error in the search for a grammatically appropriate antecedent.
Finally, we can consider the cause of the much higher acceptance of Principle B violations in studies that fail to make the grammatical deictic interpretation of the pronoun sufficiently accessible. Children in these studies who entertain the anaphoric interpretation of the pronoun should receive strong semantic support for this interpretation, given its prominence in the story, and do not have a readily available deictic interpretation that can serve to inhibit the anaphoric interpretation. This might account for the acceptance rates of 40-80% observed in these studies.

Summarizing, evidence from many different studies with children indicates that 4-year-olds show good knowledge of the disjoint reference requirements imposed by Principles B and C, but that children are more prone to error in Principle B contexts. We suggest that this difference may reflect an independently motivated contrast in the search for pronoun antecedents that has been observed in on-line studies with adults. Whereas Principle C appears to act as a constraint on the generation of representations, Principle B may sometimes act as a filter on representations that are at least temporarily generated. For a child in a TVJT, who may reasonably assume that the puppet is commenting on something that could plausibly have occurred in the story, this amounts to the claim that for children Principle B acts as a constraint on selection among possible interpretations, whose effects are most apparent when a grammatical interpretation of a pronoun is accessible. This conclusion about Principle B in children may be seen as a more articulated version of an account presented by Grimshaw & Rosen (1990).

5.4 Pronouns vs. Clitics

We must also consider the frequently reported finding that the DPBE is much weaker in languages with clitic pronouns. Although there are some studies of clitic languages where children’s improved performance may simply reflect an experimental design that better satisfies the Condition of Plausible Dissent (e.g., Varlokosta 2000), there are other studies that show strong cross-language differences using the same tasks (e.g., McKee 1992), suggesting that the impact of clitic pronouns on DPBE is genuine. Therefore, it appears that English children’s pronoun interpretations are more sensitive to contextual information than children who speak clitic languages. Although McKee’s stories did not provide an accessible deictic antecedent, the Italian children rarely resorted to the anaphoric interpretation (15% acceptance). Instead, they were able to access a deictic antecedent, a strategy apparently unavailable to English-speaking children (82% acceptance of anaphoric interpretations). What accounts for this difference? Here we suggest that the interaction of pronoun type with cases of accidental coreference may be relevant. English pronouns can be used in the examples of local accidental coreference that escape ‘Rule I’, such as (13) above. In Italian and other clitic languages such examples require full pronouns and disallow clitic pronouns. This difference between clitic and full pronouns may impact the way in which children (and adults) access and inhibit potential antecedents during language comprehension.

6. Conclusion

The relation between grammatical knowledge and linguistic behavior is complex. In any experimental task, participants must access their linguistic knowledge in real time and relate it to

---

16 Nevertheless, Padilla (1990) reports 30-40% choice of anaphoric interpretations of Spanish clitic pronouns in an act-out task, and studies in Dutch using weak and strong pronouns have not reported consistent differences (see Table 3). Therefore, the empirical record is not yet unequivocal.
a host of nonlinguistic properties of the experimental context. Given this complexity, behavior in a linguistic experiment (and, for that matter, in the real world) may be determined by (a) the grammar, (b) the parser, (c) pragmatic influences on the interpretation of the context and (d) world knowledge. Thus, in order to assess whether a given behavior is reflective of grammatical structure, the experimenter must take great care to neutralize any possible influence from extra-grammatical factors. We have argued that prior findings showing the Delay of Principle B Effect (DPBE) and the Quantifiational Asymmetry (QA) leave room for extra-grammatical explanation. Indeed, once these extra-grammatical factors are removed, preschoolers show little evidence of deficit in their knowledge of Principle B.

Our argument was based on two kinds of data. First, a review of the existing literature indicates that the empirical support for the DBPE and QA is rather less robust than it is often presumed to be. We find substantial variability across experiments in the degree to which children show either effect, and we find corresponding variability in the experimental control of extra-linguistic factors. This variability serves to undermine our confidence in the conventional wisdom. Second, we conducted three new experiments in which we found no evidence for either DBPE or QA. The only time that we did find a QA is when we deliberately introduced extra-linguistic factors that we considered as possible causes of a spurious QA in previous studies. In short, children in our experimental manipulations behaved in a way consistent with knowledge of Principle B, for quantificational and referential antecedents alike.

The dissolution of the QA also resolves an apparent conflict in previous experimental literature. Whereas there are widespread reports of 4-5 year olds making errors in Principle B contexts, we also find many results that indicate that children of the same age show clear mastery of Principle C. This asymmetry between Principle B and Principle C introduced a problem for the pragmatic explanation of the QA. The standard account of the QA consists of two parts. First, the claim there are two mechanisms governing the interpretation of pronouns in Principle B contexts, one dealing with syntactic variable binding, the other dealing with pragmatic conditions on the appropriate use of pronouns. Second, the claim that children’s difficulties in Principle B contexts with referential antecedents derive from difficulty in applying the relevant pragmatic rule. The problem with this kind of explanation is that the very same pragmatic rule is assumed to apply in Principle C contexts as well. Thus, if children’s difficulties with Principle B were due to difficulty in employing this pragmatic rule, we would expect to find the same difficulty with Principle C, contrary to fact. However, to the extent that our results show that 4-5 year olds do, in fact, control Principle B, then we have eliminated the asymmetry between Principles B and C, and hence any problems that go along with it.

Nevertheless, there appears to be some basis for the asymmetry between Principles B and C, as our review of previous literature shows. In studies that present apparently fair tests, it is common for children to allow interpretations that violate Principle B 10-15% more often than interpretations that violate Principle C; in studies with inappropriate contextual controls the illicit interpretations are accepted at even higher rates. We have suggested that the reason why children are often fooled into allowing impossible antecedents in Principled B contexts but not in Principle C contexts derives from independently motivated differences in how these constraints are processed in real time. A number of adult psycholinguistic studies indicate that for pronouns in Principle B contexts, both licit and illicit antecedents are at least temporarily considered. Principle B, in this sense acts as a filter on a set of existing potential antecedents. However, in Principle C contexts, the adult psycholinguistic evidence suggests that NPs in the c-command domain of the pronoun are never even considered as possible antecedents. Principle C, thus,
applies in real time by placing anticipatory constraints on the construction of the set of possible antecedents. Thus, we argue that whatever asymmetries there are between Principle B and Principle C in children’s behavior derive from their implementation in real time. As we have shown, under appropriate experimental conditions, children show behavior consistent with adult-like knowledge of Principle B.

Acknowledgments

We are grateful to Paul Elbourne, Nina Hyams and many members of the Maryland CNL Lab for useful discussion of the issues in this paper. This work was supported in part by NSF awards #BCS-0196004 to CP and #BCS-0412809 to JL, by NIH award DC-006829 to JL, and by a fellowship from the Heiwa Nakajima Foundation to ET.

References


Lee, Ming-Wei, and John N. Williams. 2006. The role of grammatical constraints in intra-sentential pronoun resolution. Ms., London Metropolitan University and Cambridge University.


Appendix A

List of target sentences used in the 8 experimental stories. Each participant saw all 8 stories, paired with the quantifiational (Q) or referential (R) target sentence, in a Latin Square design.

<table>
<thead>
<tr>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>R 2</td>
<td>Q 3</td>
</tr>
<tr>
<td>I think that every space guy decorated him.</td>
<td>I think that every Space Guy decorated his costume.</td>
<td>I think that every astronaut decorated him.</td>
</tr>
<tr>
<td>R 1</td>
<td>Q 2</td>
<td>R 3</td>
</tr>
<tr>
<td>I think that Storm trooper decorated him.</td>
<td>I think that Storm Trooper decorated his costume.</td>
<td>I think that Alien decorated him.</td>
</tr>
<tr>
<td>Q 2</td>
<td>R 3</td>
<td>Q 4</td>
</tr>
<tr>
<td>I think that every superhero squirted him.</td>
<td>I think that every Superhero squirted his body.</td>
<td>I think that every knight squirted him.</td>
</tr>
<tr>
<td>R 2</td>
<td>Q 4</td>
<td>R 5</td>
</tr>
<tr>
<td>I think that Robocop squirted him.</td>
<td>I think that Robocop squirted his body.</td>
<td>I think that Dog squirted him.</td>
</tr>
<tr>
<td>Q 3</td>
<td>R 4</td>
<td>Q 5</td>
</tr>
<tr>
<td>I think that every lizard sprayed him.</td>
<td>I think that every lizard sprayed his body.</td>
<td>I think that every lizard sprayed him.</td>
</tr>
<tr>
<td>R 3</td>
<td>Q 5</td>
<td>R 6</td>
</tr>
<tr>
<td>I think that the Blue Lizard sprayed him.</td>
<td>I think that Blue Lizard sprayed his body.</td>
<td>I think that Butterfly sprayed him.</td>
</tr>
<tr>
<td>Q 4</td>
<td>R 5</td>
<td>Q 6</td>
</tr>
<tr>
<td>I think every smurf stamped him.</td>
<td>I think every smurf stamped his shirt.</td>
<td>I think every smurf stamped him.</td>
</tr>
<tr>
<td>R 4</td>
<td>Q 6</td>
<td>R 7</td>
</tr>
<tr>
<td>I think that Painting Smurf stamped him</td>
<td>I think Painting Smurf stamped his shirt.</td>
<td>I think Dog stamped him.</td>
</tr>
<tr>
<td>Q 5</td>
<td>R 6</td>
<td>Q 7</td>
</tr>
<tr>
<td>I think that every dwarf painted him.</td>
<td>I think that every dwarf painted his costume.</td>
<td>I think that every dwarf painted him.</td>
</tr>
<tr>
<td>R 5</td>
<td>Q 7</td>
<td>R 8</td>
</tr>
<tr>
<td>I think that Grumpy painted him.</td>
<td>I think that Grumpy painted his costume.</td>
<td>I think that Smurf painted him.</td>
</tr>
<tr>
<td>Q 6</td>
<td>R 7</td>
<td>Q 8</td>
</tr>
<tr>
<td>I think that every troll labeled him.</td>
<td>I think that every Troll labeled his shirt.</td>
<td>I think that every troll labeled him.</td>
</tr>
<tr>
<td>R 6</td>
<td>Q 8</td>
<td>R 1</td>
</tr>
<tr>
<td>I think that Orange Troll labeled him.</td>
<td>I think that orange Troll labeled his shirt.</td>
<td>I think that grey bear labeled him.</td>
</tr>
<tr>
<td>Q 7</td>
<td>R 8</td>
<td>Q 2</td>
</tr>
<tr>
<td>I think that every turtle wiped him.</td>
<td>I think that every turtle wiped his shirt.</td>
<td>I think that every turtle wiped him.</td>
</tr>
<tr>
<td>R 7</td>
<td>Q 2</td>
<td>R 3</td>
</tr>
<tr>
<td>I think that Blue turtle wiped him.</td>
<td>I think that blue turtle wiped his shirt.</td>
<td>I think that Mickey wiped him.</td>
</tr>
<tr>
<td>Q 8</td>
<td>R 3</td>
<td>Q 1</td>
</tr>
<tr>
<td>I think that every M&amp;M fanned him.</td>
<td>I think that every M&amp;M fanned his shirt.</td>
<td>I think that every M&amp;M fanned him.</td>
</tr>
<tr>
<td>R 8</td>
<td>Q 1</td>
<td>R 4</td>
</tr>
<tr>
<td>I think that Hat M&amp;M fanned him.</td>
<td>I think that Hat M&amp;M fanned his body.</td>
<td>I think that Barney fanned him.</td>
</tr>
</tbody>
</table>