Revised 1/18/08

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

Anastasia Conroy[1], Eri Takahashi[1], Jeffrey Lidz[1,2], Colin Phillips[1,2]

[1] Department of Linguistics, [2] Neuroscience and Cognitive Science Program
University of Maryland

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 (Elbourne, 2005). Here we attempt to resolve the status of the QA with a series of 3 studies using the Truth Value Judgment Task and a review of over 30 previous studies. Using an improved experimental design, we show that children disallow local pronoun binding with both 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. Drawing on evidence from adult language processing, we suggest that Principle B acts as a filter on representations rather than as a constraint on structure generation. This filter, when combined with pragmatic infelicities in the tasks used, may account for the wide variability in the strength of the DPBE found in previous studies.

1. Introduction

It is rare that data from child language are taken to constrain models of adult grammatical competence. One such case concerns what Elbourne (2005) calls the *Quantificational Asymmetry* (QA) in children's application of Principle B, a constraint that prohibits 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 *Mama Bear washed herself*. The observation that, with some types of antecedents, young children are delayed in demonstrating knowledge of Principle B has been termed the Delay of Principle B Effect (DPBE). However, when the referential subject in (1) is replaced with a quantificational subject, as in (2), children no longer allow the corresponding anaphoric interpretation. The observation that children display knowledge of Principle B with quantificational, but not referential antecedents has been termed the Quantificational Asymmetry (QA).

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

The theoretical argument, due originally to Chien and Wexler (1990), concerns the scope of Principle B in the grammar. Chomsky's (1981) binding theory treats all cases of anaphora as involving coindexation, thus positing no difference between the treatment of quantificational and referential antecedents with respect to binding. However, Reinhart's (1983) approach distinguishes bound variable anaphora from other cases of coreference. Consequently, the quantificational asymmetry in children's interpretations is taken to mirror the asymmetry that is independently posited in Reinhart's binding theory. In this case, the results from child language acquisition have been taken as strong evidence for a distinction between bound variables and coindexation in the grammar.

The Quantificational Asymmetry has been an important finding in the study of child language because it appears to decide among leading views of anaphora, lending crucial support to Reinhart's theory. However, Elbourne (2005) has raised a number of concerns regarding the validity of the QA, although without providing supporting developmental evidence. In this article we investigate children's knowledge of Principle B, particularly with respect to the methodological concerns raised by Elbourne. We suggest that many of the prior observations concerning the Quantificational Asymmetry and DPBE reflect shortcomings of the experimental tests used, and do not reflect properties of children's grammars. We present a series of three novel experiments designed to examine children's knowledge of Principle B, addressing Elbourne's methodological concerns. We find that appropriately controlled experiments appear to eliminate the QA, a finding that partly vindicates Elbourne's claims. However, we find that 4-year old children overwhelmingly respect Principle B in sentences with referential and quantificational antecedents alike, disconfirming Elbourne's prediction that children would show a DPBE for all antecedents.

We claim that the QA is an artifact of experimentation, and suggest that the DPBE is less pervasive than is standardly reported. Children do, in fact, obey Principle B. 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 (for recent reviews see Büring 2005, Elbourne 2007).

Interestingly, our findings align Principle B more closely with other constraints, such as Principle C, a constraint on backwards anaphora that rules out coreference in sentences like (3). This constraint has repeatedly been shown to be mastered quite early (Crain & McKee 1985, Crain & Thornton 1998, Guasti & Chierchia 1999/2000, Kazanina & Phillips 2001; Leddon & Lidz 2005).

(3) She washed Mama Bear.

We conclude that there is no QA, and that at the level of grammatical representation there is no DPBE. However, we are left with an open question. If it is the case that children adhere to Principle B, a grammatical constraint that bans coreference between a pronoun and a local antecedent, then why are children highly susceptible to errors under certain experimental conditions? One might expect that if children's grammars disallow the representation associated with a Principle B violation, then even biased methodologies should not cause them to access interpretations that violate Principle B. We argue that children's greater susceptibility to Principle B errors (over Principle C errors) derives from independently motivated properties of anaphoric dependency processing that are revealed in adult on-line studies on Principles B and C. This argument, in concert with our findings from children, removes a potential impediment to grammatical 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 regarding 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 three experiments that address the methodological concerns raised concerning previous experiments. In these experiments we find no evidence for a QA or for a DPBE. These new results leave us with the question of why findings about children's knowledge of Principle B are so discrepant. In Section 5 and an appendix we review over 30 previous studies on binding constraints in children. We 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 on-line 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.

In one reading the pronoun refers to a sentence external antecedent. However, even if we require that the pronoun be anaphoric to the subject of the sentence, an ambiguity remains. The ambiguity can be seen more clearly when we place such sentences in a VP-ellipsis context, as in (5) (Keenan 1971, Sag 1976, Williams 1977).

(5) Al loves his sister and Bill does too
  a. = Al_i loves his_i sister and Bill_j loves his_j sister too ('sloppy' reading)
  b. = 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*. This simple paradigm tells us 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 interpretation 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. = every linguist$_i$ loves his$_i$ sister and Bill$_j$ loves his$_j$ sister too (sloppy)
  b. ≠ every linguist$_i$ loves his$_i$ sister and Bill$_j$ loves his$_i$ 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, examples like those 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 (Keenan 1971, 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. In particular, a bound variable must be c-commanded by its antecedent (7a) (but cf. Büring 2005, Elbourne 2007), whereas no such restriction holds for accidental coreference (7b).

(7) a. The people who work for him$_{*i/j}$ love every department chair$_i$.
  b. The people who work for him$_{i/j}$ 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.

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 observation that the disjoint reference constraints (i.e., Principles B and C; Chomsky 1981) and bound variable anaphora apply only to c-command relations. 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_i likes $him_{*i/j}$
b. Al_i's sister likes $him_{i/j}$

(10) a. $He_{*i/j}$ thinks that I like Al_i
b. $His_{i/j}$ sister thinks that I like Al_i

This parallel in the domain of applicability is expected if the binding constraints apply only to bound variable anaphora.

The idea that binding constraints apply only to bound variable anaphora correctly 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 $candidate_i$ likes $him_{*i/j}$.
(12) Al_i likes $him_{*i/j}$.

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 recognized that her version of Principle B is insufficient to block coreference in (12), due to the possibility of accidental coreference.[1] She proposed that this binding theoretic loophole is closed by an additional constraint, labeled Rule I (14).

(14) *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 rule 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) λx. x 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.

[1] 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).

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) λx. x 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 A Developmental Dissociation

While the theoretical arguments in favor of restricting Principle B to cases of bound variable anaphora can stand on their own, evidence from the developmental pattern of adherence to Principle B is often presented as the best possible evidence for the existence of two different mechanisms underlying 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 example, in a classic picture judgment study (Chien & Wexler 1990) 5-year olds accepted a coreferential intepretation 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 claims that different mechanisms restrict quantificational and referential antecedents for pronouns, and the dissociation observed in the child data comports well with this claim. 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% to 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 as well with Reinhart's theory of binding as is sometimes suggested. 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. Consider the dialog in (17):

(17) A. Is that John?
B. It must be. He's wearing John's coat.

Here speakers do not detect a Principle C violation, despite the fact that *he* and *John* are taken to be identical in reference. Rule I is satisfied, however, since an assertion that *someone is wearing John's coat* is distinct from an assertion that *someone is wearing his own coat*. The fact that these assertions can be distinguished licenses coreference.

[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.

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 & Thornton 1998), because it is a task that encourages the experimenter to make very explicit his assumptions about the discourse context against which a sentence is judged. Although many of the issues discussed below generalize to other experimental measures, and we believe that the concerns that we raise encompass a variety of tasks, we are not able to examine all tasks with the level of detail that we apply to the TJVT. We next turn to a discussion of the basic components of a TVJT test of constraints on anaphora.

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 (18) and (19), 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 regarding whether these should be handled by one or two mechanisms in the grammar.

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

Clearly, we cannot ask young children to give explicit judgments about the range of coreference possibilities for such sentences. The Truth Value Judgment Task was thus devised to tackle the problem of probing complex grammatical phenomena in young children using a task that is engaging and that 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. After the story, Kermit makes a statement about the story and the child's task is to judge whether Kermit's statement was accurate. The experimenter can use the child's judgment to draw inferences about the child's interpretation of the target sentence. The task has many advantages. It can be used to probe complex grammatical representations in 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 investigating 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 allowing experimenters to manipulate the discourse context and control for its potential impact 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 (18) or (19) is both true and prominent in the story, but does not judge (18) or (19) to be true statements about the story, then the child presumably did not access the anaphoric interpretation of the pronoun. Assuming that the task has been executed properly, we infer that the reason why the child did not access the anaphoric interpretation is because a binding constraint made that interpretation inaccessible.

However, the simple judgment that the TVJT requires of a child also presents its greatest challenges. The child makes a judgment about a sentence and a story, and then the experimenter must make an inference concerning the child's grammar. Because the setup of the story is integral to the range of interpretations available to the child, great care must be taken to satisfy the assumptions underlying the task, so as to avoid misleading results (Crain & Thornton 1998).

A widespread assumption in TVJT studies is that children will assent to the truth of a sentence if they can ('Principle of Charity'). In the case of a potentially ambiguous sentence, a test sentence is typically presented in a scenario that makes only one reading of the sentence true. Thus, if a child assents to the truth of the sentence in a scenario, we presume that the child has access to the interpretation made true in that scenario. If the child rejects the sentence in a scenario, then this rejection is taken as evidence that the interpretation made true in that scenario is unavailable. However, such responses only justify conclusions about the child's grammar if we can rule out extra-linguistic reasons for the responses. If an interpretation is too strongly biased in the scenario provided, then a child may say 'yes' due to contextual coercion of an ungrammatical interpretation. Conversely, if an interpretation is not made sufficiently available in the story, then it is possible that the child's rejection results not from grammatical constraints, but rather from properties of the discourse context. Consequently, if we are to reason about a child's grammar from yes/no responses in a TVJT, then we must ensure that the interpretations under investigation are made equally available in the experimental contexts.

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 (18) or (19) 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 focused 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. However, this depends on the assumption that the anaphoric interpretation was sufficiently available in the context that only the child's grammar could be responsible for his rejection.

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 amount to the need to balance the relative accessibility of the interpretations under investigation. Here it is important to distinguish two notions of accessibility of pronoun interpretations. One applies to the potential referents of pronouns, the other to the propositions that the pronouns appear in.[3]

Assumption 1: Availability. Pronouns are used to pick out referents that are independently available in a discourse. Therefore, in order to test whether a child's grammar allows or disallows a particular antecedent for a pronoun it is important that the intended referent for the pronoun be available in the current discourse. This requirement applies equally to antecedents of anaphoric and deictic pronouns, but in tests of binding constraints it is particularly relevant to deictic antecedents, since they need not be explicitly mentioned in the same sentence as the pronoun and hence may be overlooked. If a child is presented with a sentence containing a

[3] We would like to thank an anonymous reviewer for helping to clarify these issues.

pronoun for which his grammar excludes an anaphoric interpretation, but the context fails to make a deictic interpretation available, then the child may be 'coerced' into choosing the anaphoric antecedent because that is the only discourse-accessible antecedent.

Relatedly, it is important to establish that potential anaphoric antecedents are considered to be potential antecedents by children, once the possible effects of binding constraints 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 (Roeper 1985, Koster 1994), then children may avoid the anaphoric reading of (19) due to this general bias, independent of Principle B.

Assumption 2: Disputability. The second requirement involves the propositions in which pronouns are used in tests of binding constraints. In order to be submitted for a natural true/false judgment it is important that a proposition should be 'under consideration' in the experimental setting. If an interpretation has never been under consideration, then children may have difficulty rejecting that interpretation, even if it refers to an event that did not occur. For example, if a sentence like (19) is presented in a context where the narrative focus is on whether *the dwarves will all paint another individual* (deictic interpretation), and where the possibility that *each of the dwarves might paint himself* is never a live option in the story (anaphoric interpretation), then children are likely to judge the truth of the deictic interpretation, irrespective of the impact of binding constraints.

In a TVJT test of binding constraints children are typically asked to judge sentences containing pronouns following stories in which an anaphoric interpretation of the pronoun is true and a deictic interpretation is false. Practically, this entails that the story should make the deictic interpretation a genuine potential outcome at some point in the story. 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 (Crain & Thornton 1998). Previous studies of children's adherence to Principle B 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. The requirement is commonly satisfied by making the deictic interpretation of the sentence almost become true in the story, although we argue that this

is insufficient if the event that almost happens is incidental to the rest of the narrative (see also Hulsey et al. 2004).

In tests of the QA there is one further consideration that must be taken into account: the scenarios 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. This means that the child's basis for rejecting the test sentences should be maximally similar in the referential and quantificational conditions. The reasoning that we follow in the new studies described below is that if the very same event makes the test sentence false in the referential and quantificational conditions, then the risk of a spurious QA is reduced.

In addition to constraints on the set-up of a TVJ scenario, it also important that researchers are confident in their measures of the interpretation accessed by children. For this reason, children's yes/no answers should ideally be followed with requests for explanation. Although most descriptions of the TVJT focus on using children's yes/no answers to draw inferences about their grammatical representations, the real interest is in how children interpret the test sentences, for which the yes/no answers are but one convenient measure. Children's justifications for their answers provide important additional information on their interpretations, and they sometimes indicate that the yes/no answer is misleading. For example, if a child answers 'no', but his explanation indicates that he interpreted the pronoun anaphorically, then his 'no' answer clearly provides no evidence for avoidance of anaphoric interpretations.

Elbourne's (2005) critique of prior studies of the DPBE gives primary attention to one aspect of how these assumptions are satisfied, emphasizing factors such as differential availability of the anaphoric antecedent that 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. Issues of disputability, and in particular the question of whether the (ultimately false) deictic interpretation of the test sentence was under consideration in the scenario, are less of a focus in Elbourne's critique, but we suggest that inadequate satisfaction of this assumption may be at least as important as failures to satisfy the availability assumption in tests of Principle B and the QA in children. We show in Experiment 3 that when the disputability

assumption is not properly satisfied in the referential condition, the appearance of an asymmetry arises.

3.2 An Example: Thornton & Wexler (1999)

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 the 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 and the Appendix.

The story in (20) 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 (21) or (22).

(20) 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 of himself. Bert thanks the helpful reindeer for starting to brush him. He say's he's sorry he can't reciprocate by helping brush the helpful reindeer; he needs to finish brushing all the snow off of himself because he's still very cold.

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

The sentences in (21) and (22) are both true in the scenario in (20) 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 (21) as true in 58% of trials and accepted (22) as true in only 8% of trials. These results suggest a strong DPBE for referential antecedents and a clear QA.

The scenario in (20) meets some basic desiderata of a test of binding constraints in children, in the respect that the anaphoric interpretation of the pronouns in (21-22) is true in the story and there is a deictic interpretation of the pronoun that is false in each case. Furthermore, the Condition of Plausible Dissent is at least partly met in the quantificational and referential conditions. However, when we consider more systematically how the availability and disputability assumptions are satisfied in the two conditions we find contrasts that may account for the observed QA, independent of Principle B.

The availability assumption is differentially satisfied in the quantificational and referential conditions, due to the contrasting accessibility of the anaphoric and deictic antecedents assumed in each condition. In the referential condition (21) the anaphoric antecedent is *Bert* and the relevant deictic antecedent is *the third reindeer*, whose help Bert briefly considers reciprocating. In the quantificational condition (22), on the other hand, the anaphoric antecedent is *every reindeer* and the relevant deictic antecedent is *Bert*. Because Bert is clearly the main protagonist of the story, and because he is the anaphoric antecedent in the referential condition and the deictic antecedent in the quantificational condition, it is perhaps not surprising that children's judgments were based on the anaphoric interpretation on most referential trials and the deictic interpretation on most quantificational trials. Thus the QA can be derived by assuming that children associate the pronoun *him* with the most prominent referent in the story, with no need to appeal to Principle B.

Elbourne (2005) points to the concern that the main protagonist (Bert) plays a different role in the two conditions, and also raises the concern that children may have favored the deictic interpretation in the quantificational condition (22), perhaps due to a general dispreference for bound variable interpretations of pronouns, an assumption that we directly test in Experiment 2.[4]

Turning to the disputability assumption, we again find that the referential and quantificational conditions differ in how the test sentence relates to the central theme of the story

[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.

(who will brush Bert?), and do so in a way that could explain children's contrasting responses in the two conditions. In the referential condition the anaphoric interpretation corresponds to the proposition that *Bert brushed himself*, something that is clearly true and is closely related to the focus of the story on how Bert can remove the snow from his body. In contrast, the deictic interpretation corresponds to the proposition that *Bert brushed the third reindeer*, a possibility that Bert mentions only in passing and that is not directly related to the theme of the story. This may have reduced the accessibility of the deictic interpretation in the referential condition and contributed to children's bias to judge the anaphoric interpretation. In the quantificational condition the anaphoric interpretation corresponds to the proposition that *every reindeer brushed himself*, something that is clearly true and that is at least indirectly related to the theme of the story, since the reindeer refused to help Bert because they were busy helping themselves. The deictic interpretation corresponds to the proposition that *every reindeer brushed Bert*. Although this proposition does not come close to becoming true, it is closely related to the theme of the story, since Bert does ask each reindeer in turn to help him. Thus there is an asymmetry in how the test sentence relates to the theme of the story across the two conditions, making the disputability assumption unsatisfied by the overall design. The deictic interpretation is more relevant to the theme of the story in the quantificational condition than in the referential condition, and so it is perhaps not surprising that children judged the deictic interpretation more frequently in the quantificational condition than in the referential condition.

In light of these concerns, and others raised by the survey of previous studies discussed in Section 5 below, we conducted a series of three experiments that were designed to provide a fairer test of QA and DPBE.

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 a 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 features that raised concerns 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 a pronoun direct object NP, and either a referential or quantificational subject NP. 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 with respect to the scenario.

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 items that were intermixed with the target items to create a test consisting of 16 stories. Filler stories were included to provide an independent measure of the children's understanding of the task. Furthermore, the filler sentences were assigned dynamically, such that the experimenter provided either a true or false target sentence, in order to balance the overall

number of true and false sentences presented over the course of the experiment. For individual children the stories were divided over 2 sessions of no more than 20 minutes, with each session 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 a sample story in (23) and then review how the story satisfies 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. The text in (23) describes a scenario presented to children, and does not represent the child-friendly narrative that the children actually heard. Sample videos and slides illustrating the stories can be found on the authors' web sites.

(23) *The Painting Story*[5]
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

[5] Note that 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, paralleling earlier studies. Note also that although a number of the characters in the stories carried out reflexive actions (e.g., painting themselves), explicit reflexives were never used in the telling of the stories.

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 Lead-in: 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 …
Quantificational Lead-in: 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 …

(24)	Grumpy painted him.	*Referential condition*
(25)	Every dwarf painted him.	*Quantificational condition*

The story in (23) 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.

The referential and quantificational conditions are closely matched, both in terms of the accessibility of antecedents/referents (availability assumption) and in terms of the accessibility of the propositions that children were asked to judge (disputability assumption). The central character in the story is Hiking Smurf, and he is the intended deictic referent in both the referential and quantificational conditions. This minimizes any potential concerns about the availability of a suitable deictic referent for the pronoun. Meanwhile, the anaphoric antecedents are matched as closely as possible in the two conditions. Grumpy is the anaphoric antecedent in the referential condition. He is a prominent character in the story and is associated with the most vivid event in the narrative. Also, he is the most prominent of the set of dwarves who are the anaphoric antecedent in the quantificational condition. Hence, the availability assumption is satisfied in a similar fashion in the two conditions.

Turning to the disputability assumption, the referential and quantificational conditions are again closely matched and all relevant propositions are live possibilities during the story. The

anaphoric interpretation corresponds to the proposition that *Grumpy painted himself* (referential condition) and that *every dwarf painted himself* (quantificational condition). Both propositions are true, and it is exactly the same event that makes both propositions true. After the first two dwarves paint themselves Grumpy shows great reluctance to paint himself, temporarily raising the possibility that the anaphoric interpretations will fail to become true in both conditions. When Grumpy finally agrees to paint himself, both propositions become true, and in a similarly vivid fashion. Meanwhile, the deictic interpretation corresponds to the proposition that *Grumpy painted Hiking Smurf* (referential condition) and that *every dwarf painted Hiking Smurf* (quantificational condition). Both propositions are false, and for exactly the same reason, i.e., Grumpy's refusal to help.

Furthermore, in response to Elbourne's concern that children may have difficulty in accessing the bound variable interpretation of singular pronouns we took further steps to promote the accessibility of the anaphoric interpretation in the quantificational condition. First, all characters in the story have a clear individual identity in addition to being part of a group. This contrasts with the relatively undifferentiated reindeer in the story in (20). Second, each character draws attention to his need to paint himself, before offering assistance to Hiking Smurf. (Note, however, that no explicit reflexives were ever used in the telling of the stories.) Third, both Hiking Smurf and Grumpy/the dwarves are mentioned in the lead-in sentence that precedes the test sentence' (this last measure is similar to Thornton &Wexler 1999).

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. If the design succeeds in matching the conditions in terms of the accessibility of the relevant referents and propositions then any observed differences between conditions can more confidently be attributed to the children's grammars. Of course, 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 more direct test of the accessibility of the anaphoric interpretation.

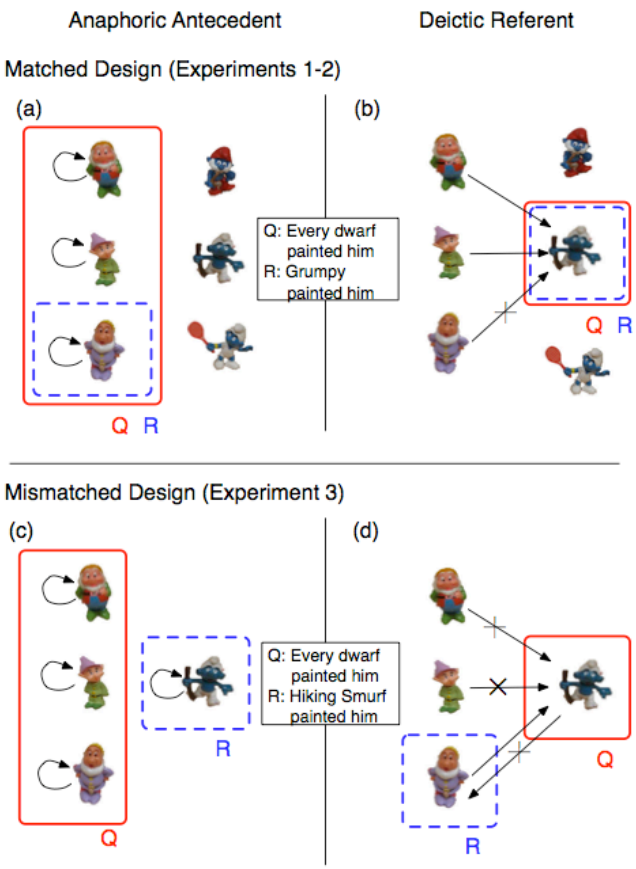

Figure 1: Schematic illustration of characters and scenarios used in Experiments 1-3, highlighting the events that made the anaphoric interpretation of the pronoun true (left column) and the deictic interpretation of the pronoun false (right column), and comparing the referential condition (dotted blue 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).

4.1.2 Results

Results are based on the number of trials in which the responses reflected an anaphoric interpretation of the pronoun, which was always true in the story, and the number of trials where the 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 the children were always asked to explain to Kermit why he was right or wrong, under the guise of helping him to do a better job. In cases where a child's justification conflicted with his yes/no judgment the justification 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. This procedure is consistent with the basic logic of the TVJT (Crain & Thornton 1998). In Experiment 1 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 anaphoric 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 made the test sentence false. The choice of deictic interpretation was consistent across conditions. Assuming that the logic of the TVJT was adequately satisfied, this result suggests that the children and adults avoided 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 anaphoric 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 compatible 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. (26) and (27) have identical truth conditions to the illicit anaphoric interpretation of the sentences in (24) and (25), 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. Nevertheless, there was still no 'correct' or 'incorrect' response in this Experiment, since the test sentences were fully ambiguous.

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

Note that the sentences in (26) and (27) 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). Because 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 evidence for anaphoric 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 $Z = -0.836$; $p = 0.40$). The 13 rejections in the referential condition were contributed by 13 children, such that no child rejected the target sentence in the referential condition more than once. The 17 rejections in the quantificational condition were contributed by 13 children, 4 of whom gave two rejections. Adults accepted the anaphoric 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 × antecedent interaction. 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 χ^2 = 42.395, p = 0.001).

4.2.3 Discussion

Experiment 2 was identical to Experiment 1 in all respects, except for the position of the pronoun in the target statements. In Experiment 1, Principle B was potentially active, but by making the pronoun a possessor in Experiment 2 we neutralized any possible contribution of Principle B. 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 anaphoric 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, contrary to concerns raised in some previous studies (e.g., Koster 1994, Elbourne 2005). 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 difference is proportionally very small compared to the differences that have been presented as evidence for the QA.[6]

In addition, the results of Experiment 2 may provide independent support for the Principle of Charity. In this study both the deictic and the anaphoric interpretations of the test sentence were available and grammatical, but only one of them was true in the story. The results showed that the children overwhelmingly said that the sentence was true, which corroborates the basic assumption of the TVJT that children show a bias to give positive answers (Crain & Thornton 1998).

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 confirm Elbourne's prediction that appropriately matched referential and quantificational

[6] The difference in the current study was a 23% higher rate of 'yes' answers in the referential condition than in the quantificational condition. This is very different from the proportional increases attributed to a QA in previous studies (e.g., Chien & Wexler 1990: 220%; Thornton & Wexler 1999: 625%; Matsuoka 1997: 250%; Philip & Coopmans 1996: 160%).

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 frequently cited argument in favor of Reinhart's approach, although 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, as 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 Principles B and C that has been seen in previous studies. 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.[7]

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 statistical 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 review 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.

[7] 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.

4.3 Experiment 3

4.3.1 Design and Participants

In Experiments 1-2 we showed that children appear to respect Principle B and show no Quantificational Asymmetry when presented with tests that satisfy the logic of the TVJT. In addition, we argued in Section 3 that the QA observed in previous studies may derive from experimental designs that failed to equally satisfy the availability and disputability assumptions in referential and quantificational conditions. We therefore predict that it should be possible to reintroduce the contrasts in children's behavior by altering key features of our experimental designs. Accordingly, in Experiment 3 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 (20).

The story in (28) is a modified version of the story in (23), and the test sentences are shown in (29-30). 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 differences that might emerge in children's responses. Rather, it serves as an initial test of how much the contextual details of a TVJT experiment influence children's interpretation of pronouns. As in Experiments 1-2, the text in (28) presents the plot of a story told to children, and does not give the exact narrative that the children heard. Example videos are available from the authors' web sites.

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

Referential Lead-in: This was a story about dwarves and Hiking Smurf.[8]
Quantificational Lead-in: This was a story about dwarves and Hiking Smurf.

(29) Hiking Smurf painted him. *Referential condition*
(30) Every dwarf painted him. *Quantificational condition*

The story in (28) and the test sentences in (29-30) differ from their counterparts in Experiments 1 and 2 by reintroducing the contrasts between the referential and quantificational conditions found in the story in (20) and the test sentences in (21-22).

In terms of the accessibility of antecedents/referents for the pronoun *him* (availability assumption) the referential and quantificational conditions differ. The central figure in the narrative is Hiking Smurf, who fulfills different roles in the two test sentences: he is the intended anaphoric antecedent in the referential condition and is the intended deictic antecedent in the quantificational condition. Therefore, if children simply interpret *him* as referring to Hiking Smurf they should judge the test sentence true in the referential condition and false in the quantificational condition, leading to the appearance of a QA. Meanwhile, there is an additional motivation for children to judge the ungrammatical anaphoric interpretation of the referential condition, since the intended deictic referent in that condition is the relatively insignificant and undifferentiated third dwarf.

In terms of the accessibility of relevant propositions (disputability assumption) the referential and quantificational conditions again contrast, and in a way that could lead to a spurious QA. In the referential condition the anaphoric interpretation of the pronoun is associated with the proposition that Hiking Smurf painted himself. This is clearly true in the story, although it was not an expected outcome. The deictic interpretation of the pronoun is associated with the proposition that *Hiking Smurf painted the third dwarf*, an eventuality that is unrelated to the plot of the story and is never under consideration until Hiking Smurf mentions in passing that he cannot do it. Thus there is a likely bias in the referential condition for the anaphoric interpretation. The situation is reversed in the quantificational condition. The deictic interpretation is associated with the proposition that *every dwarf painted Hiking Smurf*. Although

[8] As in Experiments 1-2 the lead-in sentences first mentioned the appropriate (deictic) antecedent and then the inappropriate (anaphoric) antecedent before presenting the test sentence. However, the exact format of the lead-in sentences was changed in order to match our understanding of Thornton & Wexler's procedure.

this proposition does not become true in the story, it is directly related to the central theme of the story, i.e., Hiking Smurf's request to every dwarf for help. The anaphoric interpretation corresponds to the proposition that every dwarf painted himself, something that is clearly true in the story, although it is indirectly related to Hiking Smurf's quest.

It should be noted that the story in (28) nominally conforms 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 and introduces events that potentially make the grammatical intepretation plausible. But as we have tried to emphasize, a fair test of QA and DPBE 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 to each child across two testing sessions. Participants were an additional 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

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 statistically reliable (Wilcoxon Signed Ranks Z = -2.507; p = 0.01). The total of 36 anaphoric interpretations of referential trials were contributed by 22 children, 14 of whom gave two non-adultlike responses. In the quantificational condition, the total of 10 anaphoric interpretations were contributed by 7 children, 3 of whom gave two non-adultlike responses. 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.

% Accept Binding/Coreference	*Experiment 1*		*Experiment 2*		*Experiment 3*	
Children						
Referential Antecedent	7/64	11%	51/64	80%	36/64	56%
Quantificational Antecedent	9/64	14%	47/64	73%	10/64	16%
Adults						
Referential Antecedent	3/64	5%	53/64	83%		
Quantificational Antecedent	2/64	3%	43/64	67%		

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 exact cause of the different results in Experiment 1 and Experiment 3, because 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 in the referential condition. We counted the number of trials on which children justified their answers by referring to the events in the story that falsify the target sentence. In the quantificational condition children gave a relevant rationale on 32 of the 54 trials where they answered 'no', 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, the relevant events were referred to on only 1 of the 28 trials where a 'no' answer was given. That is, almost no children said that the sentence was false because Hiking Smurf was too busy to help the third dwarf. The difference in justification-type between conditions likely reflects the contrasting accessibility of the referents and propositions associated with the deictic interpretation in the two conditions.

In the referential condition of this experiment there were a particularly high number of trials in which children gave 'no' responses but gave a justification that was odd in two respects. First, these justifications were not consistent with the 'no' response. Second, unlike in experiment 1, these justifications did not make reference to the event in the story that falsified the target sentence on the noncoreferential interpretation.

With respect to the first point, it appeared to us that these justifications were most consistent with an anaphoric interpretation of the pronoun. For example in 30% (19/64) of trials children responded to *Hiking Smurf painted him* by saying something like "He [first dwarf] didn't have

enough time, he [second dwarf] wouldn't, then he [third dwarf] painted him". This justification consists of three conjoined sentences with identically interpeted VPs. In each of these VPs, it is obvious that the object of the verb refers to Hiking Smurf. Under the assumption that the justification reflects the child's interpretation of the target sentence, it follows that the child also interpreted the pronoun in the target sentence as referring to Hiking Smurf. This leads us to believe that these justifications reflect an anaphoric interpretation of the target sentence, and so were scored in that fashion. Of course, this reasoning is valid only to the extent that the children's justifications reflect their interpretation of the target sentence, as is standardly assumed in these types of tasks (Crain & Thornton 1998). However, if this assumption does not hold, we have no independent confirmation that the children are giving yes/no responses based on their interpretation of the sentence relative to the context.

Returning to the second point, if these 'no' responses reflect genuine noncoreferential interpretations, these justifications are incongruent in a different respect. In Experiment 1, a typical response for a noncoreferential interpretation described the event in the story that falsifies the target sentence. In response to a sentence like *Hiking Smurf painted him*, where the grammatical antecedent is the dwarf, children typically responded by saying, "No, because he [the smurf] didn't have time [to paint the dwarf]." This response serves two functions. First, by making reference to the event in the story that falsifies the grammatical interpretation, this response illustrates that children interpreted the pronoun in an adult-like fashion. Second, the justification elaborates coherently on their yes/no response, giving us more confidence in our result. These two features of children's justifications are criterial for taking their judgments to reflect their understanding of the sentence-context pair. It is striking that in Experiment 3, responses that make reference to the falsification event are essentially missing, as noted above. Because the yes/no responses and justifications were not aligned, we chose to score this type of response on the basis of the most coherent intepretation of the justification with respect to the target sentence.

Of course, the 'no' responses that we are taking to reflect an anaphoric interpretation are still 'no' responses, a fact that highlights the oddity of the experimental context in Experiment 3. That is, the misalignment between the yes/no response and the justification can itself be taken as

additional evidence that the contexts in which these sentences were presented were somehow unnatural. Such misalignments were found only in the referential condition of Experiment 3 and not in any of the other 5 conditions that we tested, supporting our contention that apparent Principle B violations in the previous literature might reflect imperfectly designed experimental materials and not a lack of knowledge on the part of children.[9]

One further piece of evidence suggests that the children had difficulty relating the test sentences to the story. In some trials children responded simply by retelling the story, as if they were not sure 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.

Thus far we have shown that when presented with suitably balanced experimental conditions children appear to abide by the disjoint reference constraints imposed by Principle B, and show no evidence of a Quantificational Asymmetry. We characterized the prerequisites for a fair test of children's knowledge in terms of accessibility of antecedents/referents (availability) and relevant propositions (disputability), and we showed in Experiment 3 that when we reintroduced contrasts in how these requirements are satisfied, based on scenarios used by Thornton & Wexler (1999) and others, both a QA and DPBE reemerged. Nevertheless, we should note that our experiments do not allow us to conclusively establish the relative importance of availability and disputability, since our focus is on the improvements that obtain when both are satisfied.

Although our experiments lend support to the notion that children know Principle B (and associated constraints, such as Rule I), our findings are at odds with the received wisdom on this topic. Therefore, we next turn to a survey of previous studies on DPBE and QA in order to determine whether the received wisdom is consistent with what has been found, and whether variation in previous results can be understood in terms of the experimental design factors that we have identified here.

[9] One thing that remains mysterious is that when our child participants were confused by the referential condition, they answered 'no', whereas the prior literature has found a higher rate of 'yes' responses in such conditions. We suggest that this difference may be explained by other properties of the experimental sessions, possibly by the nature of the fillers. In our task, the truth of the filler sentences was determined dynamically to ensure a balance of 'yes' and 'no' responses across the experimental session, whereas other studies (e.g., Thornton & Wexler 1999) fixed the truth of the filler sentences independently of the child's responses to be the opposite of the grammatical response. Since the grammatical response in previous work was associated with a 'no' response, then this means that all fillers had "yes" responses, possibly giving rise to an overall 'yes' bias.

5. Previous Findings

We have shown that under appropriate experimental conditions children abide by Principle B, showing no effects of a QA and no DPBE, contrary to the received wisdom on this topic. In this section we survey more than 30 previous studies and find that there is a wide divergence of findings across studies, and that 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. Finally, we face the remaining question of why children, if they have knowledge of Principle B, are so susceptible to interpretations that violate this constraint but not to violations of other constraints on anaphora. We suggest that this contrast may be related to a parallel contrast between Principles B and C found in recent studies on the real-time processing of anaphora. Here we summarize the overall findings from the survey and some comments that are relevant to a number of previous studies. Further discussion of the specifics of certain individual studies can be found in Appendix B.

5.1 The Quantificational Asymmetry

We examined 19 studies that tested for a Quantificational Asymmetry, among which 10 report a QA and 9 do not.[10] These studies are summarized in Table 2. Among the studies that report a QA the rates of acceptance in the referential condition varies from 31% to 93%, and the rates of acceptance in the quantificational condition vary from 0% to 27%. Among the studies that do not find a QA some find that children perform well with referential and quantificational

[10] Our survey 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, 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.

antecedents alike (Kaufman 1988, Hestvik & Philip 1999, current studies) whereas others find that children show similarly high error rates for both types of antecedent (Lombardi & Sarma 1989, Boster 1991 Exp 2, Avrutin & Wexler 1992, Utakis 1995, Grolla 2005). Thus, there is substantial variability in the results of previous studies (see also Kaufman 1994, Koster 1994, Elbourne 2005 for earlier reviews), and the finding of a QA is certainly not consistent.

Among the studies in our survey, Kaufman (1988) is one of the earliest tests of the QA, and it 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. It is therefore striking that Kaufman reports almost identical rates of acceptance of around 16% for the two conditions, thereby providing evidence against both the QA and the DPBE. One other study shows very low error rates for both types of antecedent. 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. The authors 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.

Another group of studies shows no QA, while finding similarly high rates of errors in referential and quantificational conditions. Many of the errors can be attributed to limitations of the experimental designs used. For example, Grolla (2005) used a picture verification task in which only an anaphoric antecedent was provided in the test scenarios, leaving children with no alternative candidate referent for the pronoun. Avrutin & Wexler (1992) conducted a TVJT study on Russian that shares many design features with the scenario in (20) that we have discussed at length in Section 3. It is therefore not surprising that this study showed error rates in the referential condition that were similar to those reported by Thornton & Wexler (1999), although

it is surprising that the Russian children showed similarly high error rates in the quantificational condition. Lombardi & Sarma (1989) showed high rates of non-adultlike responses in an act-out task, which we discuss further in Appendix B.

A number of the studies that report a QA used TVJTs or similar designs that are subject to the same methodological concerns discussed in Section 3. Matsuoka (1997) used story formats that closely parallel the example in (20) from Thornton & Wexler (1999), and hence it is unsurprising that Matsuoka found a similar QA. A pair of studies by McDaniel and colleagues found a QA in tasks that are described as grammaticality judgment tasks, although they are very similar to tasks that are described elsewhere as TVJTs (McDaniel, Cairns, & Hsu 1990, McDaniel & Maxfield, 1992). Given the similarity with TVJTs, it is very relevant that these tasks did not make the deictic reading of the target sentences accessible. This could be responsible for the high rates of non-adultlike responses in the referential conditions, although it is not clear why a QA should have arisen in this study. In Appendix B we discuss further TVJT-like studies that have reported a QA, and offer specific suggestions about the source of the observed asymmetries in those studies (Thornton 1990, Boster 1991, Avrutin & Thornton 1994, Savarese 1999).

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). We contend that picture-based tasks are subject to the very same constraints that we have discussed for TVJTs, except that it is more difficult in picture-based tasks to assess how well the constraints are satisfied. In a picture verification task, as in a TVJT, children are placed in a situation where they could choose to interpret a pronoun either deictically or anaphorically. As in a TVJT, the choice of whether to interpret the pronoun deictically or anaphorically may depend on a number of factors in addition to the child's grammar, including the accessibility of suitable deictic antecedents and expectations about what events are likely to be commented upon. The primary difference between a TVJT and a picture judgment task is that in a TVJT the experimenter uses the narrative to explicitly control availability and disputability, whereas in a picture-based task a greater burden is placed on the child to conjure up a relevant context in which to interpret the picture. Elbourne (2005) offers a number of suggestions about

how a QA may have arisen in Chien & Wexler's classic study, in which children were shown a line drawing and told "Here is Goldilocks, and here are the three bears. Is every bear touching her?" We refer the reader to Elbourne's paper for specific comments on that study.

Author	Language	Age	*N*	Accept Referential	Accept Quant.	Method
Studies reporting no Quantificational Asymmetry						
Kaufman 1988	English	2;7-3;11	30	23%	18%	TVJT
		5;0-6;5	30	10%	13%	
Lombardi & Sarma 1989	English	4;0-6;2	11	55%	49%	Act out, TVJT
Boster 1991, Exp. 2	English	3;3-4;9	24	38%	42%	Picture Verification
Avrutin & Wexler 1992	Russian	4-7	16	52%	41%	TVJT
Utakis 1995	English	3;4-9;5	30	37%	40%	TVJT
Baauw, Escobar & Philip 1997	Spanish; clitic	mean 5;6	45	10%	10%	Picture Verification
Hamann, Kowalski & Philip 1997	French; clitic	3;5-4;8	9	22%	30%	Picture Verification
		5;3-5;11	8	0%	12%	
Hestvik & Philip 1999	Norwegian	4;5-5;11	15	9%	3%	Picture Verification
Grolla 2005	English	3;7-5;11	23	52%	46%	Picture Verification
	Braz. Portug.	3;4-6;6	40	44%	49%	
Studies reporting a Quantificational Asymmetry						
Chien & Wexler 1990, Exp. 4	English	2;6-3;11	48	70%	54%	Picture Verification
		4;0-4;11	45	60%	40%	
		5;0-5;11	44	51%	16%	
		6;0-7;0	40	24%	14%	
McDaniel, Cairns & Hsu 1990	English	2;9-6;7	19	44%	19%	Grammaticality Judgment
Thornton 1990	English; "who"	3;7-4;8	12	49%	8%	TVJT
Boster 1991, Exp. 1	English; "who"	4;6-6;0	10	38%	4%	TVJT
McDaniel & Maxfield 1992	English	3;1-6;10	37	41%	25%	Grammaticality Judgment
Avrutin & Thornton 1994	English; collective vs. distributive	3;10-4;10	33	93%	27%	TVJT
Philip & Coopmans 1996	English	3;6-7;0	19	68%	26%	Picture Verification
	Dutch; strong pronoun	4;3-6;11	37	66%	50%	
Matsuoka 1997	English	3;10-6;0	19	70%	20%	TVJT
Savarese 1999	English	3;5-5;11	25	31%	N/A	TVJT
		4;3-6;1	26	N/A	0%	
Thornton & Wexler 1999	English	4;0-5;1	19	58%	8%	TVJT

Table 2: Summary of results from tests of the Quantificational Asymmetry in children.

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. Our criticisms apply to specific studies, and not to the task itself.

Summarizing, our studies are by no means unique in failing to find a QA, and most previous studies that have reported a QA are amenable to alternative explanations that do not invoke a grammatical asymmetry between coreference and variable binding. We therefore consider it well justified to doubt the received wisdom about the existence of a QA in children.

5.2 The Delay of Principle B Effect

Our experimental 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. In addition to the 19 studies that tested the QA, our survey included a further 14 studies that tested children's adherence to Principle B with referential antecedents only. The results of these studies are summarized in Table 3. Even if we restrict our attention initially to studies on English, we find 13 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%.

The high degree of variability in acceptance rates across studies suggests that children's responses are not simply the product of a relatively stable grammar. If each study had 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 reflects specifics of the tasks used in individual studies. Indeed, in our survey we found that a good deal of the variability in previous results could be explained by task differences, and in particular by the extent to which the task provided a clear deictic alternative to the illicit anaphoric

interpretation of the pronoun.[11] Nevertheless, we find that there is a 'residue' of the DPBE that is a real effect and not an experimental artifact, and we propose an account for this effect in Section 5.3.

In tests of DPBE where the QA is not at stake it remains important to satisfy the assumptions of availability and disputability for the anaphoric and deictic interpretations of the pronoun. However, it is more straightforward to satisfy these assumptions, because there is no need to also closely match quantificational and referential conditions. In all TVJT tests of DPBE that we are aware of the anaphoric interpretation of the pronoun is made true and readily accessible. The primary variation in the experimental designs lies in whether a deictic antecedent is readily available, and in whether the proposition corresponding to the deictic interpretation of the pronoun is a live possibility in the scenario, despite ultimately turning out to be false.

In TVJT studies where the deictic interpretation of the pronoun is accessible in the context, we find relatively low rates of non-adultlike judgments, although the rates are often too high to be dismissed as experimental 'noise' (Kaufman 1988, 16% acceptance[12]; Thornton 1990, 29% acceptance across conditions; Boster 1991, 21% acceptance across conditions; Savarese 1999, 31% acceptance; Kiguchi & Thornton 2004, 27% acceptance; current study, 11% acceptance). In these studies the deictic interpretation of the pronoun corresponds to a prominent character in the story, and the event corresponding to the deictic interpretation is explicitly rejected in the story.

In a number of other studies using TVJT or similar tasks we find that the experimental design either fails to make a deictic antecedent available, or provides a deictic interpretation that is not seriously under consideration in the scenario. In these studies we typically find much higher rates of acceptance of illicit anaphoric interpretations of pronouns (Grimshaw & Rosen 1990, 42% acceptance; McDaniel, Cairns, & Hsu 1990, 44% acceptance; McDaniel & Maxfield 1992, 41% acceptance; McKee 1992, 82% acceptance; Matsuoka 1997, 70% acceptance; Thornton & Wexler 1999, 58% acceptance; current study, exp. 3, 56% anaphoric interpretations). We offer more specific remarks on some of these studies in Appendix B, but the overall generalization

[11] We also found substantial variation across studies in the level of detail provided in the experimental descriptions. In many cases there was insufficient information to allow an adequate assessment of the methods used.

[12] 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.

from TVJT studies of DPBE is clear: children are more likely to give an ungrammatical anaphoric interpretation in tasks where a grammatical deictic interpretation of the pronoun is not readily accessible.

Study	Language	Age	*N*	Accept Ref.	Task
Jakubowicz (1984)	English	4 5	10 11	30%† 25%†	Act out
Wexler & Chien (1985)	English	2;6-6;6	129	43%	Picture Selection
Deutsch, Koster & Koster (1986)	Dutch; strong pron.	6	32	46%	Picture Selection
Solan (1987)	English	4-7	37	57%	Act Out
Chien & Wexler (1990), Exps 1-2	English	2;6-6;6	298	29%†	Act Out
Grimshaw & Rosen (1990)	English	4-5	12	42%	TVJT
Padilla (1990)	Spanish; clitic	3;0-3;11 5;0-5;11	20 20	37% 30%	Act out
McKee (1992)	English Italian; clitic	2;6-5;3 3;7-5;5	60 30	82% 15%	TVJT TVJT
Sigurjónsdóttir & Hyams (1992)	Icelandic	4;0-4;6 4;6-5;0	10 10	45% 43%	TVJT
Baauw (1999)	Dutch; weak pron.	4;2-5;3	15	47%	Picture Verification
Varlokosta (2000)	Greek; clitic Greek; strong pronoun	3;7-5;6	20	5% 13%	TVJT
Varlokosta & Dullaart (2001)	Greek; clitic Greek; strong pron. Dutch; weak pron. Dutch; strong pron.	3;3-7;6	10	5% 5% 55% 45%	TVJT
Kiguchi & Thornton (2004)	English	4;1-5;10	13	27%	TVJT
Spenader et al. (2007)	Dutch	4;4-6;6	83	25%†	Picture Verification

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; the average for Spenader et al. (2007) is a mean across three sentential contexts.

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 an act-out task conducted with around 300 children. Rather higher rates of anaphoric interpretations have been reported in other act-out studies, but in at least one of these cases the design may have led to exaggeration of the number of anaphoric interpretations (Lombardi & Sarma 1989; see Appendix B).

A number of studies of DPBE have used picture verification tasks, and these have revealed similarly broad variability in acceptance of illicit anaphoric interpretations of pronouns, ranging from 9% to 68% (Chien & Wexler 1990 Exp. 4, Boster 1991, Philip & Coopmans 1996; Hestvik & Philip 1999; Spenader et al. 2007). As discussed above, picture verification tasks rely on a similar logic to TVJTs, with the difference that it is harder to control the context against which the child judges the test sentence. We refer the reader to Elbourne (2005) and to Spenader et al. (2007) for discussion of factors that may impact how children choose to interpret pronouns in these tasks. A variant on the picture verification task is the picture selection task used in an important early study by Wexler & Chien (1985). In that study 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 corresponded to Principle B violations on 43% of trials. The similarity between this task and picture verification tasks depends on how a child chooses to undertake the task. The child may treat the task as a series of picture verification tasks, looking at each picture in succession and deciding whether the test sentence accurately describes the picture. In this case exactly the same considerations apply as in TVJTs and picture verification tasks. On the other hand, if the child carries out the task by first constructing an interpretation of the test sentence, and then looking for a picture that matches that interpretation, then the task is a little different. This way of approaching the task may make requirements such as disputability or plausible dissent irrelevant, but it does not remove the need for a viable deictic antecedent for the pronoun. Due to this uncertainty and the limited information about the

materials from this study we cannot offer firm suggestions on the cause of children's non-adultlike responses. However, we speculate that it may not have been immediately apparent to children that the possessor *Cinderella* was a viable referent for the pronoun (see Spenader et al. 2007 for remarks on the application to Principle B studies of Centering Theory, Grosz, Joshi, & Weinstein 1995).

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.

However, we cannot conclude from the survey 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. If Principle B acted as a strong constraint on children's interpretations, then we should expect it to be sufficiently powerful to make children 'blind' to illicit anaphoric interpretations of pronouns, something that appears not to be the case. In addition, we must also acknowledge that some studies in other languages, particularly languages with clitic pronouns, have shown that similar or identical tests elicit much lower rates of Principle B violations. 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 survey 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 as 'noise'. 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).[13] 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 error rates 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 (and 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 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 hence it is difficult to dismiss children's success in tests of Principle C as an experimental artifact. 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 found 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

[13] 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.

antecedents for a pronoun before excluding them from consideration. 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. 2007), whereas the results for Principle B are 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 & Swinney 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 greater difficulty than adults in inhibiting and recovering from incorrect initial interpretations of sentences (e.g., Hamburger & Crain 1984, Trueswell et al. 1999). Therefore, what appears in adults as transient effects of ungrammatical antecedents might appear in children as ungrammatical interpretations that persist.

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. 2007). 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 reference 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 (intrasentential or extrasentential) 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 significantly increased acceptance of Principle B violations in the studies that fail to make a grammatical deictic interpretation of the pronoun sufficiently accessible. Children in these studies who entertain the anaphoric interpretation of the pronoun receive strong semantic support for that interpretation, given its prominence in the story, and they do not have a readily available deictic interpretation that can 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 (see also 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 disputability assumption (e.g. Varlokosta 2000), there are other studies that show strong cross-language differences using the same tasks, suggesting that the impact of clitic pronouns on

DPBE is genuine.[14] For example, McKee (1992) reported that the Italian children in her study rarely accepted anaphoric interpretations of a clitic pronoun (15% acceptance), whereas English-speaking children accepted the anaphoric interpretation on a majority of trials (82% acceptance). Although the high acceptance rate in the English-speaking children could be attributed to the lack of an accessible deictic antecedent in McKee's stories, this cannot explain the strong cross-language difference. Here we suggest that the cross-language difference may be related to the availability of cases of accidental coreference. 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 tonic pronouns and disallow clitic pronouns. This difference between clitic and tonic 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 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 behavior is reflective of grammatical structure, an 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 Quantificational Asymmetry (QA) leave room for extra-grammatical explanation, and that once 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 survey of the existing literature indicates that the empirical support for the DBPE and QA is less robust than it is often presumed

[14] 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.

to be. There is substantial variability across experiments in the evidence for either effect, and corresponding variability in the experimental control over extra-linguistic factors. Second, we conducted three experiments in which we found no evidence for either DBPE or QA. However, we did find a QA when we deliberately introduced extra-linguistic factors that we considered as possible causes of a spurious QA in previous studies. Children in the first two experiments 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 same pragmatic rule is assumed to apply in Principle C contexts as well, and thus we should expect to find the same difficulty with Principle C, contrary to fact. To the extent that our results have eliminated the developmental asymmetry between Principles B and C they also eliminate theoretical problems that the contrast created.

Nevertheless, there appears to be some residual basis for the contrast between Principles B and C, as our survey of previous studies shows. Although we have argued that experimental design factors may account for much of the variability found in previous studies of DPBE, we cannot overlook the fact that children appear to be more susceptible to interpretations that violate Principle B than they are to interpretations that violate Principle C. We have suggested that this contrast derives from independently motivated differences in how these constraints impact real time language processing. Evidence from adult psycholinguistic studies suggests that NPs in the c-command domain of a pronoun are never even considered as possible antecedents, such that

comprehenders are effectively 'blind' to interpretations that violate Principle C. In contrast, a number of studies of pronouns in Principle B contexts indicate that both licit and illicit antecedents are at least temporarily considered. Children's susceptibility to Principle B violations is compatible with knowledge of Principle B, just as it is for adults.

Acknowledgments

We are grateful to two anonymous reviewers, Paul Elbourne, Nina Hyams and many members of the Maryland CNL Lab for valuable feedback on 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

Avrutin, Sergey. 1994. *Psycholinguistic investigations in the theory of reference.* Doctoral dissertation, MIT, Cambridge, Mass.

Avrutin, Sergey, and Rosalind Thornton. 1994. Distributivity and binding in child grammar. *Linguistic Inquiry* 25:165-171.

Avrutin, Sergey, and Kenneth Wexler. 1992. Development of Principle B in Russian: Coindexation at LF and coreference. *Language Acquisition* 2:259-306.

Baauw, Sergio. 1999. The role of the clitic-full pronoun distinction in the acquisition of pronominal coreference. In *Proceedings of the 23rd annual Boston University Conference on Language Development*, ed. by Annabel Greenhill, Heather Littlefield, and, Cheryl Tano, 32-43. Somerville, Mass.: Cascadilla Press.

Baauw, Sergio, María A. Escobar, and William Philip. 1997. A delay of Principle Beffect in Spanish speaking children: The role of lexical feature acquisition. In *Language Acquisition: Knowledge Representation and Processing: Proceedings of GALA '97,* ed. by A. Sorace, C. Heycock, R. Shillcock. HCRC, Edinburgh.

Badecker, William, and Kathleen Straub. 2002. The processing role of structural constraints in the interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28:748-769.

Boster, Carole Tenny. 1991. Children's failure to obey Principle B: Syntactic problem or lexical error? Ms., University of Connecticut, Storrs.

Büring, Daniel. 2005. *Binding theory*. Cambridge, UK: Cambridge University Press.

Chien, Yu-Chin, and Kenneth Wexler. 1990. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition* 1:225-295.

Chomsky, Noam. 1981. *Lectures on government and binding.* Dordrecht: Foris.

Chomsky, Noam. 1986. *Knowledge of Language: Its nature, origin, and use.* New York: Praeger.

Clifton, Charles, Shelia Kennison, and Jason Albrecht. 1997. Reading the words *her*, *him*, and *his*: Implications for parsing principles based on frequency and on structure. *Journal of Memory and Language* 36:276-292.

Cowart, Wayne, and Helen S. Cairns. 1987. Evidence for an anaphoric mechanism within syntactic processing: Some reference relations defy semantic and pragmatic constraints. *Memory and Cognition* 15:318-331.

Crain, Stephen, and Cecile McKee. 1985. The acquisition of structural restrictions on anaphora. In *Proceedings of NELS 16,* ed. by Stephen Berman, Jaewoong Choe, and Joyce McDonough, 94-110. Amherst: University of Massachusetts, GLSA.

Crain, Stephen, and Cecile McKee. 1987. Children's understanding of coreference: a pragmatic vs. a structural explanation. Paper presented at the LSA Annual Meeting, San Francisco, CA.

Crain, Stephen, and Rosalind Thornton. 1998. *Investigations in Universal Grammar: A guide to experiments on the acquisition of syntax and semantics.* Cambridge, Mass.: MIT Press.

Deutsch, Werner, Charlotte Koster, and Jan Koster. 1986. What can we learn from children's errors in understanding anaphora? *Linguistics* 24:203-225.

Elbourne, Paul. 2005. On the acquisition of Principle B. *Linguistic Inquiry* 36:333-365.

Elbourne, Paul. 2007. The interpretation of pronouns. *Language and Linguistics Compass* (in press).

Evans, Gareth. 1980. Pronouns. *Linguistic Inquiry* 11:337-362.

Foster-Cohen, Susan. 1994. Exploring the boundary between syntax and pragmatics: Relevance and the binding of pronouns. *Journal of Child Language* 21:237-255.

Gordon, Peter. 1996. The truth-value judgment task. In *Methods for assessing children's syntax,* ed. by Dana McDaniel, Cecile McKee, and Helen Cairns, 211-232. Cambridge, Mass.: MIT Press.

Grimshaw, Jane, and Sara Rosen. 1990. Knowledge and obedience: The developmental status of the binding theory. *Linguistic Inquiry* 21:187-222.

Grodzinsky, Yosef, and Tanya Reinhart. 1993. The innateness of binding and coreference. *Linguistic Inquiry* 24:69-101.

Grosz, Barbara, Aravind Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21:203-226.

Guasti, Maria Teresa. 2004. *Language acquisition: The growth of grammar.* Cambridge, Mass.: MIT Press.

Guasti, Maria Teresa, and Gennaro Chierchia. 1999/2000. Reconstruction in child grammar. *Language Acquisition* 8:129-170.

Hamann, Cornelia, Odette Kowalski, and William Philip. 1997. The French "delay of Principle B" effect. In *Proceedings of the 21st annual Boston University Conference on Language Development*, ed. by Elizabeth Hughes, Mary Hughes, and Annabel Greenhill, 205-219. Somerville, Mass.: Cascadilla Press.

Hamburger, Henry, and Stephen Crain. 1984. Acquisition of cognitive compiling. *Cognition* 17:85-136.

Heim, Irene. 1998. Anaphora and semantic interpretation: A reinterpretation of Reinhart's approach. In *The interpretive tract*, ed. by Uli Sauerland and Orin Percus, 205–246. MIT Working Papers in Linguistics 25. Cambridge, Mass.: MIT, Department of Linguistics and Philosophy, MITWPL.

Hestvik, Arild, and William Philip. 1999/2000. Binding and coreference in Norwegian child language. *Language Acquisition* 8:171-235.

Higginbotham, James. 1983. Logical form, binding, and nominals. *Linguistic Inquiry* 14:547-593.

Hulsey, S., V. Hacquard, D. Fox and A. Gualmini (2004). 'The Question-Answer Requirement and Scope Assignment', in *Plato's Problems: Papers on Language Acquisition. MITWPL 48*. Cambridge, MA: MITWPL.

Jakubowicz, Celia. 1984. On markedness and binding principles. In *Proceedings of NELS 14,* ed. by Charles Jones, and Peter Sells, 154-182. Amherst: University of Massachusetts, GLSA.

Kaufman, Diana. 1988. Grammatical and cognitive interactions in the study of children's knowledge of binding theory and reference relations. Doctoral dissertation, Temple University, Philadelphia, Pennsylvania.

Kaufman, Diana. 1994. Grammatical or pragmatic: Will the real Principle B please stand? In *Syntactic theory and first language acquisition: Cross-linguistic perspectives,* vol. 2, *Binding, dependencies, and learnability*, ed. by Barbara Lust, Gabriella Hermon, and Jaklin Kornfilt, 177-200. Hillsdale, N.J.: Lawrence Erlbaum.

Kazanina, Nina, Ellen Lau, Moti Lieberman, Masaya Yoshida, and Colin Phillips. (2007). The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language* 56:384-409.

Kazanina, Nina, and Colin Phillips. 2001. Coreference in child Russian: Distinguishing syntactic and discourse constraints. In *Proceedings of the 25th annual Boston University Conference on Language Development*, ed. by Anna H.-J. Do, Laura Domiínguez, and Aimee Johansen, 413-424. Somerville, Mass.: Cascadilla Press.

Keenan, Ed. 1971. Names, quantifiers, and the sloppy identity problem. *Papers in Linguistics* 4:211-232.

Kennison, Shelia. 2003. Comprehending the pronouns *her*, *him*, and *his*: implications for theories of referential processing. *Journal of Memory and Language* 49:335-352.

Kiguchi, Hirohisa, and Rosalind Thornton. 2004. Binding principles and ACD constructions in child grammars. *Syntax* 7:234-271.

Koster, Charlotte. 1994. Problems with pronoun acquisition. In *Syntactic theory and first language acquisition: Cross-linguistic perspectives*, vol. 2, *Binding, dependencies, and learnability*, ed. by Barbara Lust, Gabriella Hermon, and Jaklin Kornfilt, 201-226. Hillsdale, NJ: Erlbaum.

Lasnik, Howard. 1976. Remarks on coreference. *Linguistic Analysis* 2:1-22.

Leddon, Erin, and Jeffrey Lidz. 2006. Reconstruction effects in child language. In *Proceedings of the 30th annual Boston University Conference on Language Development*, ed. by David Bamman, Tatiana Magnitskaia, and Colleen Zaller, 328-339. Somerville, Mass.: Cascadilla Press.

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.

Levinson, Stephen. 2000. *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, Mass.: MIT Press.

Lombardi, Linda, and Jaya Sarma. 1989. Against the bound variable hypothesis of the acquisition of Principle B. Paper presented at the annual meeting of the Linguistic Society of America, Washington, D.C.

Lust, Barbara, Julie Eisele, and Reiko Mazuka. 1992. The binding theory module: Evidence from first language acquisition for Principle C. *Language* 68:333-358.

Lust, Barbara, Loveland, K, and R. Kornet. 1980. The development of anaphora in first language: Syntactic and pragmatic constraints. *Linguistic Analysis* 6:359-391.

Matsuoka, Kazumi. 1997. Binding conditions in young children's grammar: Interpretation of pronouns inside conjoined NPs. *Language Acquisition* 6:37-48.
McDaniel, Dana, Helen Cairns, and Jennifer Hsu. 1990. Binding principles in the grammars of young children. *Language Acquisition* 1:121-139.
McDaniel, Dana, and Thomas Maxfield. 1992. Principle B and contrastive stress. *Language Acquisition* 2:337-358.
McKee, Cecile. 1992. A comparison of pronouns and anaphors in Italian and English acquisition. *Language Acquisition* 2:21-54.
Nicol, Janet, and David Swinney. 1989. The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research* 18:5-19.
Padilla, Jose. 1990. *One the definition of binding domains in Spanish.* Dordrecht: Kluwer.
Philip, William, and Peter Coopmans. 1996. The double Dutch delay of Principle B effect. In *Proceedings of the 20th annual Boston University Conference on Language Development*, ed. by Andy Stringfellow, Dalia Cahana-Amitay, Elizabeth Hughes, and Andrea Zukowski, 576-587. Somerville, Mass.: Cascadilla Press.
Reinhart, Tanya. 1983. *Anaphora and semantic interpretation.* London: Croom Helm.
Reinhart, Tanya. 2006. *Interface strategies*. Cambridge, Mass: MIT Press.
Rosen, T. John, and Sara Rosen. 1994. Inferring the innateness of syntactic knowledge. In *The Proceedings of the 26th annual Child Language Research Forum,* ed. by Eve Clark, 71-81. Standford: CSLI.
Runner, Jeffrey, Rachel Sussman, and Michael Tanenhaus. 2003a. Assignment of reference to reflexives and pronouns in picture noun phrases: Evidence from eye-movements. *Cognition* 89:B1-B13.
Runner, Jeffrey, Rachel Sussman, and Michael Tanenhaus. 2003b. The influence of Binding Theory on the on-line reference resolution of pronouns. *Proceedings of the North Eastern Linguistics Society* 34.
Runner, Jeffrey, Rachel Sussman, and Michael Tanenhaus. 2006. Processing reflexives and pronouns in picture noun phrases. *Cognitive Science* 30:193-241.
Russell, Bertrand. 1948. *Human knowledge: Its scope and limits.* London: Allen and Unwin.
Sag, Ivan. 1976. Deletion and logical form. Doctoral dissertation, MIT, Cambridge, Mass. [Published 1980, New York: Garland.]
Savarese, Frederick. 1999. Studies in coreference and binding. Doctoral dissertation, University of Maryland, College Park.
Sigurjónsdóttir, Sigríður, and Nina Hyams. 1992. Reflexivization and logophoricity: Evidence from the acquisition of Icelandic. *Language Acquisition* 2:359-413.
Solan, Lawrence. 1983. *Pronominal reference: Child language and the theory of grammar*. Dordrecht: Reidel.
Solan, Lawrence. 1987. Parameter setting and the development of pronouns and reflexives. In *Parameter setting,* ed. by Thomas Roeper and Edwin Williams, 189-210. Dordrecht: Reidel.
Spenader, Jennifer, Erik-Jan Smits, and Petra Hendriks. 2007. Coherent discourse solves the pronoun interpretation problem. Ms. University of Groningen.
Sturt, Patrick, Hamutal Kreiner, and Simon Garrod. 2005. Talk presented at the U. of Maryland.
Thornton, Rosalind. 1990. Adventures in long-distance moving: The acquisition of complex *wh*-questions. Doctoral dissertation, University of Connecticut, Storrs.
Thornton, Rosalind, and Kenneth Wexler. 1999. *Principle B, VP-ellipsis, and interpretation in child grammar.* Cambridge, Mass.: MIT Press.

Trueswell, John, Irina Sekerina, Nicole Hill, and Marian Logrip. 1998. The kindergarten-path effect: studying on-line sentence processing in young children. *Cognition* 73:89-134.

Utakis, Sharon. 1995. Quantification and definiteness in child grammar. PhD dissertation, CUNY, New York.

Varlokosta, Spyridoula. 2000. Lack of clitic-pronoun distinctions in the acquisition of Principle B in child Greek. In *Proceedings of the 24th annual Boston University Conference on Language Development*, ed. by S. Catherine Howell, Sarah Fish, and Thea Keith-Lucas, 738-748. Somerville, Mass.: Cascadilla Press.

Varlokosta, Spyridoula, and Joanna Dullaart. 2001. The acquisition of pronominal reference by Greek-Dutch bilingual children: Evidence for early grammar differentiation and autonomous development in bilingual first language acquisition. In *Proceedings of the 25th annual Boston University Conference on Language Development*, ed. by Anna H.-J. Do, Laura Domínguez, and Aimee Johansen, 780-791. Somerville, Mass.: Cascadilla Press.

Wexler, Kenneth, and Yu-Chin Chien. 1985. The development of lexical anaphors and pronouns. *Papers and Reports on Child Language Development* 24: 138-149.

Williams, Edwin. 1977. Discourse and Logical Form. *Linguistic Inquir*y 8:101-139.

Appendix A: Experimental Materials

List of target sentences used in the 8 experimental stories. Each participant saw all 8 stories, paired with the quantificational (Q) or referential (R) target sentence, in a Latin Square design. Sample slides and videos illustrating the stories are available from the authors' web sites.

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

Appendix B: Discussion of Previous Experiments

In this Appendix we discuss a number of additional reports of the QA and DPBE, with comments that are relevant to the design of individual studies, rather than to all studies of children's knowledge of binding constraints.

A. Act-out study: Lombardi & Sarma (1989). Lombardi & Sarma 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 can only give a toy to the monkey if he first fetches it himself. Therefore, if a child interpreted the pronoun correctly but failed to complete the action, the response would be classified as ungrammatical. 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.[15] Therefore, it is possible that Lombardi and Sarma's task may have led to an exaggerated appearance of failure with Principle B.

B. Studies with wh-phrase antecedents. Thornton (1990) and Boster (1991, Exp. 1) compare children's judgments of sentences with a referential NP or a *wh*-phrase as a subject, as illustrated in (31-32), and use this comparison as a test of the QA.[16]

(31) Bert and Huckleberry Hound scratched them.
(32) I know who scratched them – Bert and Huckleberry Hound.

Both of these target sentences can be tested with an identical story. For example, Big Bird, Snuffleupagus, 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

[15] 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).

[16] Thornton (1990) uses the examples in (31-32) 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 examples shown here. Both Thornton and Boster included at least some trials with singular pronouns among the test sentences in their studies, e.g., *I know who dressed her – Baby Sally* (Boster 1991).

because of their special suits. Big Bird and Snuffleupagus 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 interpretation of the plural pronoun), but false that Bert and Huckleberry Hound scratched Big Bird and Snuffleupagus (deictic reading). The logic of the studies as tests of the QA relies on the assumption that the subject NP in (31) is referential, whereas the *wh*-phrase subject in (32) is quantificational. Both studies reveal a very clear contrast between the referential and *wh*-phrase conditions, and thus appear to present strong support for the QA. However, we have two concerns about these tasks. First, the presence of *who* does not ensure a bound variable interpretation for the pronoun; second, there are extra-grammatical interpretive mechanisms that may have led children to respond differently in the two conditions.

In a sentence with a referential subject like (31) 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 (32) with a *wh*-phrase subject presents exactly the same interpretive possibilities for the pronoun as the sentence in (31). The *wh*-phrase obligatorily binds the gap in subject position, but the *wh*-phrase is not directly or obligatorily linked to the object pronoun *them*. The object pronoun may be treated as a bound variable, linked to the subject, yielding the interpretation *I know who scratched themselves*, or the pronoun may be treated as referential, allowing accidental coreference, yielding the interpretation *I know who scratched Bert and Huckleberry Hound*. Since the interpretive possibilities for (31-32) 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 *who scratched himself*. 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 *who scratched Big Bird and Snuffleupagus*. Big Bird and Snuffleupagus 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 (32), 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 Snuffleupagus. Consequently, sentence (32) 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.

C. Negation and variable binding (Savarese 1999). Savarese presents a TVJT test of the QA that appears to satisfy disputability 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), but Savarese's study takes the interesting step of using the negative quantifier *no* in his quantificational condition, and the results showed a strong QA (referential - 31% acceptance; quantificational - 0% acceptance). However, the use of negation made it impossible to closely match the materials for the referential and quantificational conditions, and 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 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 these studies are related to the presence of negation, rather than to a QA.

D. Collective vs. distributive interpretations (Avrutin & Thornton 1994). Avrutin and Thornton (1994) present 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 more frequently accepted the illicit anaphoric reading of *The smurf and the clown dried them* 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 in this study may be captured without 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 uncertainty in children's judgments (e.g., Chien & Wexler 1990, Grodzinsky & Reinhart 1993). Second, when the 16 children who accepted anaphoric 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.

E. Disputability violations. Some studies of DPBE that report high acceptance rates may have 'coerced' children to accept illicit anaphoric interpretations by failing to provide a suitably accessible deictic interpretations. One such study is McKee (1992), the study with the highest rate of illicit responses in our survey (82%).[17] A sample story from this study involves 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 the 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

[17] 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. See Section 5.4 for discussion.

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.[18] Similar concerns about the availability of a deictic interpretation apply to studies 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 TVJTs that the children may have given truth judgments rather than well-formedness judgments. Another study in which the deictic interpretation is not a live possibility in the scenario is presented by Grimshaw and 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 have been 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.

[18] Interestingly, a very similar point is raised by Crain & McKee (1987, cited in Thornton 1990, p. 169).