



Laboratorium voor Experimentele Psychologie

Adults, children and hummingbirds

An investigation into the development of scalar implicatures and the variables that influence them

Katrijn Pipijn

Proefschrift aangeboden tot het verkrijgen van de
graad van Doctor in de Psychologie

Promotor: Prof. Dr. Walter Schaeken
Copromotor: Dr. Sara Verbrugge

2014

Copyright © 2014 by Katrijn Pipijn

Cover designed by Kristiaan Pipijn

Cover art "R. Pack 3 - Hummingbird" by cabezacondor, used under CC BY
(<http://creativecommons.org/licenses/by-nc-nd/3.0/>) /

Horizontal flip and cut from original

Cover art source:

<http://cabezadecondor.deviantart.com/art/R-Pack-3-Hummingbird-65759099>

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without written permission of the author.

Adults, children and hummingbirds. An investigation into the development of scalar implicatures and the variables that influence them

Katrijn Pipijn

Supervisor: Prof. Dr. Walter Schaeken; Co-supervisor: Dr. Sara Verbrugge

In the present dissertation we will make an in-depth investigation of scalar implicatures with an emphasis on developmental studies. There are a few different aspects we will focus on. In the first two studies, we will first of all look at the effects of task difficulty. In the literature the effects of load on pragmatic reasoning has been shown, but other studies have shown that there is no effect of task difficulty. We believe that the lack of this effect was due to the task paradigms that were used. Using altered versions of these paradigms we hope to show that there are in fact effects of task difficulty and that these are similar to load effects. Secondly, we will look at the effect of response measure on pragmatic reasoning. More specifically, we will look at the effect of responding on a scale, compared to giving binary responses. These effects have been studied in the literature before and we intend to replicate them using the earlier mentioned altered task paradigms. In a third study, we will look at reversed scalars, a specific item type that has, to our knowledge, not been investigated in detail in previous implicature work. The reversed scalars are an item type that has been considered a control item in previous research but we believe it is not. We will investigate how these items are processed at different ages and we will propose a theory to explain them: the violation tolerance hypothesis. In the final chapter, we will look at scalar diversity, or how reasoning is dependent on the specific scalar implicatures that are used.

In the first two studies, we found effects of task difficulty with children. These effects were not present in adults or preschoolers though. We believe that for adults, all the different tasks that we used were still too easy. With none of the tasks being difficult enough, we were not able to burden their cognitive resources enough for them to experience difficulty in producing scalar implicatures. For preschoolers we had to adjust the more difficult task to make it more child-friendly, but by doing this we probably minimized the difference in difficulty between the used tasks. We found effects of response measure, for all age groups. This clearly shows that using a scalar response option gives us much more information than the conventional two-alternative forced choice measure does. Like we expected, children interpreted reversed scalars differently than what is predicted by the literature and even adults showed a tendency to interpret them incorrectly too. For our third study, we used an experiment to test our violation tolerance hypothesis but the results of this study did not confirm it. We believe however that this was due to the fact that we tested adults and that testing children will give more conclusive evidence in one direction or the other. For the final chapter we found a large diversity in the processing of different Horn scales. Many studies on scalar implicatures have focused on the flagship example of <some, all>. This study joins a growing amount of research that addresses the issue that there is in fact large scalar diversity and not all the different scales can just be put together.

Volwassenen, kinderen en kolibries. Een onderzoek naar de ontwikkeling van scalaire implicaturen en de variabelen die ze beïnvloeden

Katrijn Pipijn

Promotor: Prof. Dr. Walter Schaeken; Copromotor: Dr. Sara Verbrugge

In deze verhandeling voeren we een diepgaand onderzoek uit naar scalaire implicaturen met een nadruk op ontwikkelingsonderzoek. Er zijn een paar verschillende aspecten waar we ons op hebben gefocust. In de eerste twee onderzoeken keken we naar de effecten van de moeilijkheid van de taak. In de literatuur werd herhaaldelijk een effect van cognitieve belasting op ons pragmatisch redeneren aangetoond, maar andere studies konden geen effect van taakmoeilijkheid vinden. We geloven dat het gebrek aan dit effect in deze studies te wijten is aan de taakparadigma's die werden gebruikt. Door de gebruikte taakparadigma's aan te passen hoopten we toch het effect van taakmoeilijkheid te kunnen aantonen en evidentie te vinden dat deze effecten vergelijkbaar zijn aan het effect van cognitieve belasting. Ten tweede keken we naar het effect van de responsmaat op ons pragmatisch redeneren. Meer specifiek keken we naar het effect van antwoorden op een schaal in plaats van binaire antwoordmogelijkheden. Dit effect is reeds in de literatuur onderzocht en we wilden dit effect verder bevestigen met de vooraf vermelde en aangepaste taakparadigma's. Ten derde keken we naar *reversed scalars*, een specifieke term voor een itemtype dat, zover wij weten, niet eerder onderzocht werd. Dit item werd in het verleden als een controle-item beschouwd maar wij geloven dat het toch unieker is. We onderzochten hoe dit specifieke item verwerkt wordt op verschillende leeftijden en we stelden een theorie voor om dit item uit te leggen: de Overtreding Tolerantie Hypothese. In het laatste hoofdstuk keken we naar scalaire diversiteit, of hoe ons redeneren afhangt van de specifieke scalaire implicatuur die gebruikt wordt.

In de eerste twee studies vonden we een effect van taakmoeilijkheid bij kinderen. Deze effecten waren niet aanwezig bij volwassenen en kleuters. We geloven dat de taken die we gebruikt hebben in onze studies te gemakkelijk waren voor volwassenen. Omwille van deze reden werden de cognitieve vermogens van de volwassenen nooit genoeg belast tot het punt waarop hun pragmatisch redeneren eronder zou lijden. Bij de kleuters hebben we de moeilijkere taak moeten aanpassen omdat hij anders té moeilijk was, hierdoor hebben we waarschijnlijk het verschil in moeilijkheid met de gemakkelijkere taak geminimaliseerd.

We vonden effecten van de gebruikte responsmaat voor alle leeftijdsgroepen. Dit toont duidelijk aan dat een scalaire responsmogelijkheid veel meer informatie geeft dan een binaire responsmogelijkheid. Zoals we verwachtten, interpreteerden kinderen *reversed scalars* op een andere manier dan de literatuur zou voorspellen. Zelfs volwassenen toonden een kleine neiging om de items ook incorrect te interpreteren. In onze derde studie hebben we een taak gebruikt die tot doel had onze Overtreding Tolerantie Hypothese te bevestigen, maar onze resultaten spraken deze hypothese tegen. We geloven echter dat dit ligt aan het feit dat we volwassenen hebben getest en dat kinderen meer geschikt waren geweest om onze hypothese te testen. In het laatste hoofdstuk vonden we een grote diversiteit tussen verschillende Horn schalen. Vele studies in het verleden zijn gebaseerd op de schaal < sommige, alle >. Onze studie sluit zich aan bij een groeiende groep literatuur die aantoont dat er eigenlijk een grote diversiteit bestaat tussen schalen en dat men niet zomaar conclusies kan formuleren op basis van dit ene voorbeeld.

Dankwoord

Uiteraard zijn er enkele mensen die ik uitgebreid wil bedanken.

Ten eerste wil ik graag mijn promotor Walter Schaeken bedanken. Dank U wel Walter om een eerste klasse promotor te zijn. Niet enkel op wetenschappelijk vlak, maar misschien zelfs meer op persoonlijk vlak was U een enorme steun. Bedankt om me de ruimte, de tijd, het luisterend oor, de motivatie en de kansen te geven die ik nodig had. Het mag gezegd worden dat dit doctoraat niet tot een (goed) einde was gekomen zonder Uw hulp.

Ik wil ook graag Sara Verbrugge bedanken om last minute nog mee in mijn team te stappen en nog vele uren naleeswerk voor haar rekening te nemen. Ik was enorm blij dat je deze inspanning nog wou doen en je frisse blik op mijn manuscripten wou delen.

Next, I would like to show my gratitude to the members of my supervisory committee: Prof. Dr. Gerrit Storms, Dr. Wim De Neys, Prof. Dr. William Van Belle and Prof. Dr. Napoleon Katsos. By providing me with helpful suggestions about my research at that point, they have aided to bring this PhD to a good end. In addition, I would also like to thank the members of my examination committee, Prof. Dr. Napoleon Katsos, Dr. Wim De Neys, and Dr. Aline Sevenants for agreeing to read my dissertation and residing in the jury for my doctoral defense.

Ook een specifieke dank aan mijn tante Maria die de lay-out van dit boekje voor haar rekening genomen heeft. Ik ben blij dat jij de vervelendste taak uit mijn hele doctoraat van mij hebt overgenomen.

Verder wil ik graag de vele collega's bedanken die mij de voorbije jaren gezelschap hebben gehouden. Het belangrijkste aspect aan een goede job zijn goede collega's en ik heb hierbij zeker geluk gehad. De vele

lunches, drinks, feestjes en andere sociale aangelegenheden maken dat ik altijd met veel plezier zal terugdenken aan mijn tijd in het PSI. Vooral Karolien en Leen wil ik extra in de bloemetjes zetten. Jullie hadden jullie handen vol met mij als bureaugenoot, maar jullie hebben dat met glans volgehouden en ik ben blij dat ik er nu twee fantastische vriendinnen aan heb overgehouden! Bedankt om mijn handje vast te houden tijdens 'the darkest night'.

Ten slotte wil ik graag mijn familie bedanken. Bedankt om er gewoon altijd te zijn. Jullie niet-aflatend vertrouwen in dat ik dit kon, of zelfs dat ik wat dan ook kan, is altijd een enorme steun geweest!

Katrijn

—
Even the darkest night will end
And the sun will rise
—

Victor Hugo

Table of contents

Chapter 1	General introduction	I
1.1.	Overview of the experimental studies	13
1.2.	Chapter 2	14
1.3.	Chapter 3	15
1.4.	Chapter 4	16
1.5.	Chapter 5	17
	References	18
Chapter 2	Children and Pragmatic Implicatures: A Test of the Pragmatic Tolerance Hypothesis with Different Tasks	23
	Abstract	24
2.1.	Introduction	25
2.2.	Experiment 1	31
	2.2.1. Material and methods	32
	2.2.2. Results	35
	2.2.3. Discussion	38
2.3.	Experiment 2	40
	2.3.1. Method	41
	2.3.1.1. Subjects	41
	2.3.1.2. Procedure	41
	2.3.1.3. Materials	42
	2.3.2. Results	43
	2.3.3. Discussion	54
2.4.	General discussion	56
	References	60

**Chapter 3 Is it Tolerance or Pragmatic Tolerance?
The Pragmatic Tolerance Hypothesis
in preschoolers 63**

Abstract 64

3.1. Introduction 66

3.2. Method 75

 3.2.1. Participants 75

 3.2.2. Procedure 76

 3.2.3. Materials 76

3.3. Results 79

3.4. Discussion 83

References 87

**Chapter 4 An investigation of the
violation tolerance hypothesis 89**

Abstract 90

4.1. Introduction 91

4.2. Method 102

4.3. Results 103

4.4. Discussion 108

References 114

Chapter 5	How scales are influenced by scales	117
Abstract		118
5.1. Introduction		119
5.2. Experiment 1		123
5.2.1. Method		126
5.2.2. Results		129
5.2.3. Discussion		132
5.3. Experiment 2		134
5.3.1. Method		136
5.3.2. Results		137
5.3.3. Discussion		140
5.4. General discussion		143
References		147
Chapter 6	Final Discussion	151
References		172

1

General introduction

In August 2013, Google introduced Google Hummingbird. Google Hummingbird was the new algorithm that supports Google search. Previously, if you would look up something on Google, the old algorithm would look at the exact words that you were searching for, or at synonyms of these words. The new Hummingbird algorithm goes much further than that. On top of the exact words, it looks at the context and the meaning of these words. Hummingbird looks at the intent of the person carrying out the search, it will look at the whole sentences and interpret them. The main focus is 'why is this person typing in this search, what does he or she want to know?'. These types of questions can be categorized in sub disciplines of linguistics and are called semantics and pragmatics. On the one hand we have semantics, the study of the meaning of words. On the other hand there is pragmatics, which studies the way in which context contributes to the meaning of these words and the sentences they are used in. Pragmatics is what we will focus on in this dissertation. Google Hummingbird is only one of the many applications of pragmatics. In our everyday communication pragmatics is used constantly too. When we speak we say a lot of things explicitly, but there might be even more things that we say implicitly. Consider the following dialogue:

- (1) - A: Can we go see a movie tonight?
- B: I have to work late.

The response given by person B does not seem to be a direct response to the question asked by A. Yet we can derive it anyway. By saying that he has to work late, person B is saying that he will not be able to go see a movie that night. On top of that, he also gives the reason why. The reason why is said explicitly and the actual response to the question was

implicit. This phenomenon cannot be explained by semantics. The person asking the question will derive his answer by using pragmatics. He will use all the information he has about the world and the other person to interpret his response and come to the right conclusion.

This is one of the spectacular potentials of the human mind. Just like Google Hummingbird, we are able to process massive amounts of context information in fractions of seconds and interpret sentences in the correct way. However, occasionally, the fascinating machinery that is our brain, fails. On the one hand, we sometimes do not interpret all the available information in the correct way, leading to an incorrect interpretation. On the other hand, this particular feature of our mind is not fully developed yet in children. For example imagine the earlier conversation occurring between a mother and child. It is easy to imagine that the child does not recognize his mother's response as an answer to his question, that he does not see the relevance of the response. Research has shown that pragmatics is particularly hard for people with Autism Spectrum Disorder (Baltaxe, 1977; De Villiers, Stainton, & Szatmari, 2007; Surian, 1996). These difficulties are present in both children and adults with Autism Disorder and Asperger syndrome. Asperger Syndrome is an Autism Spectrum Disorder (ASD) that is characterized by high-functioning children and adults that do not have significant delays or difficulties in language or cognitive development as opposed to other forms of ASD like High-functioning Autism. Although both groups have difficulty with pragmatics, adults with Asperger syndrome do seem better at deriving scalar implicatures than a matched group with High-functioning Autism (Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts 2009). It seems that those specific language skills that adults with Asperger Syndrome have but adults with High-functioning Autism do not have, are indispensable to

make pragmatic inferences. To understand why these irregularities occur, and thus how children or adults with Autism Spectrum Disorder think, but also to advance technological applications and artificial intelligence, it is important to investigate the way our mind handles pragmatics. Especially a focus on developmental research is essential to help children and people with ASD in the future. By understanding pragmatics we can influence, intervene and remediate people and especially children when necessary. Moreover, if we move towards a future in which communicating with computers and robots is an indispensable part of our daily lives, being able to explain this discipline is essential, and Google Hummingbird is one step closer towards this goal.

Of course Google Hummingbird does not only rely on one subfield of linguistics in its algorithm. There are many subfields within linguistics, for example phonology, morphology, syntax, semantics and pragmatics. Most of these will be of some importance in the algorithm. Each subfield focuses on one specific aspect of how we use words and language. In this dissertation however, we will only focus on one aspect: pragmatics. Pragmatics studies the relationship between what is said and the context in which it is said. It comes from the assumption that when we utter a sentence, more often than not, there is a lot more information to be derived if one knows the context in which the sentence is uttered. For instance it might be significant who exactly utters the sentence. For example when you are driving in a car and your passenger says they need a restroom, you know you probably will have time enough to drive to the next gas station. However, when your passenger is not an adult but a two-year-old that is in the midst of toilet training, it might be wiser to stop right away instead of driving another 10 miles. The rest of the conversation in which a sentence is said can be important as well to un-

derstand the correct meaning of an utterance. For example 'The sun is coming out' will have a different meaning when it is preceded by 'I really should get out of bed and start my day' than when it is preceded by 'winter is finally over'. In the former it means that the sun is appearing above the horizon while in the latter the sun could be interpreted more symbolic as 'a warmer season'. But not only context is important, common knowledge of linguistics is important too. When someone says 'John and Julie got married and went on vacation', most people will suppose it happened in that order; partly because most people plan a honeymoon after their wedding, but also because of the way the sentence was constructed. If the sentence was constructed another way 'John and Julie went on vacation and got married', it suddenly becomes a lot less clear in which order they had those activities planned. Yet in most social situations, the intentions of the speaker are clear to the listener and the ability to understand these intentions is called pragmatic competence. Within pragmatics, there are several areas of interest that have been researched over the years. In this dissertation we will focus on implicatures. Implicatures are pieces of information that can be derived from an utterance, but are not explicitly expressed or strictly implied. An example of an implicature is the previously mentioned sentence about John and Julie. Most people will assume that they got married before their honeymoon, so they will make the implicature. However, an important aspect of implicatures is that they are cancellable. If someone adds 'not necessarily in that order', then the implicature is cancelled but the meaning of the original sentence is not altered.

A lot of research has been done on implicatures but experimental research on the subject has really lifted off after a study by Noveck (2001). Before this time, most of the work done was theoretical work

that relied mainly on pragmatic intuitions and only in rare cases were they accompanied by observational data. The theoretical framework around implicatures is much older than the study by Noveck and started off with the work done by Grice (Grice, 1975; Grice, 1991). Grice states that people follow a cooperative principle in a conversation, which furthers the purpose of that conversation. The cooperative principle states that people follow certain rules when they participate in a conversation. These rules will adapt to the specific needs of that conversation. Grice describes these rules in what he calls Maxims. These Maxims are: Maxim of Quality, Maxim of Quantity, Maxim of Relation and Maxim of Manner (see Table 1). The Maxim of Quality states that you should only say things that you believe are true and that you should not say anything for which you lack any adequate evidence. The Maxim of Quantity states that you should limit your contribution to a conversation and only give that information that is required and nothing more. The Maxim of Relation expects you to only give information that is relevant to the conversation. Lastly, the Maxim of Manner imposes you to avoid obscurity of expression, avoid ambiguity, be brief and be orderly. The cooperative principle works in the two directions of the conversations. The speaker ought to stick to these principles in what he says, and the listener in his turn will assume that the speaker is sticking to them to interpret what is said. The interpretation of implicatures can then be explained by these Maxims. We will use our earlier example (1) about going to the movies to illustrate this. When we look at the response given by B by itself, it does not seem to be a response to the question asked by A. However, if A assumes that person B is conversing according to the cooperative principle, he will assume that the information that person B is giving, is relevant (Maxim of Relation) and that it is true (Maxim of Quality). He

can also assume that person B was trying to be brief (Maxim of Manner) and that he therefore left out a step of his train of thought. Person A will combine all this information, together with his general knowledge that it is not possible to go to the movies and work late at the same time. As a result, person A can only come to the conclusion that person B will not be able to go to the movies with him.

Table 1. Grice's Maxims of communication

Linguistic Principle	Criterion	Violations
Quality	Truth	Exaggeration, fantastical, excessive description
Quantity	Informativeness	Redundancy, repetition, excessive brevity
Relation	Relevance	Digression
Manner	Clarity	Vagueness, obliqueness, metaphor

Grice distinguishes between two types of implicatures: conventional and conversational. A conventional implicature does not rely on these maxims of conversation, because the implicature is not context dependent. Instead it is related to the form of the expression and to the conventional meaning of the words used in the sentences. Consider the following example: She is small but strong. Using the word 'but' in this sentence indicates that there is a contradiction in the sentence, more precisely between the words 'small' and 'strong'. The speaker indicates that he believes these two characteristics don't usually co-occur; but it is not explicitly said. He uses a conventional implicature to express this

opinion. Conversational implicatures on the other hand are dependent on the context. To interpret them correctly, one must take into account the meaning of the words and think according to the Maxims of communication. The examples of these conversational implicatures are described in the previous paragraphs.

In this dissertation we will focus on scalar implicatures. These are a specific type of conversational implicatures that contain a scalar term. The implicature lies in the assumption that when someone uses a less informative scalar term (for example the word *some*) in a conversation or a sentence, the use of this word implicates the negation of a stronger word with a similar meaning (for example the word *all*). This assumption can be made on the basis of the maxims described above. If the stronger word were actually more suitable, then a speaker would have used it. However, the speaker did not use it; therefore, it is probably not adequate for the situation. When a person makes the implicature and adopts the interpretation of the scalar term that was described just now, the person is reasoning pragmatically. However, a person could also interpret the scalar term more broadly, in a way that includes the stronger term instead of rejecting it. This would be a logical interpretation of the scalar term.

An important aspect of scalar implicatures is that they are cancellable (Grice, 1991). Because there is an ambiguity between the two interpretations it is possible that when the listener interprets the scalar term pragmatically, he makes some assumptions that are not correct. It is however possible that more context information becomes available later on, that contradicts the assumptions made by the listener. In this case, the meaning of the scalar term can easily be changed to the broader interpretation that entails the stronger term, or the speaker can change

his or hers initial utterance all together to an utterance with the strong scalar term. The scalar implicature will therefore be cancelled. This cancellability aspect is considered to be a good criterion as to whether or not a scalar implicature is a conversational implicature (Grice, 1991). There is a consensus in the literature that the cancellation of a scalar term in an utterance is not sufficient to speak of a conversational implicature. Whether or not the cancellability is a necessary criterion to speak of a conversational implicature however is not so clear. While Grice believes it is, both Weiner (2006) and van Kuppevelt (1996) do not believe it is necessary. Weiner has shown various cases in which cancelling an implicatures does not lead the listener to adjust his belief in the pragmatic interpretation. One of the main cases he discusses are instances of sarcasm in which the cancellation even strengthens the belief in the pragmatic interpretation. Borge (2009) argues that in this case the utterance is reinforced instead of cancelled. Therefore Borge believes that sarcastic utterances are special cases and that cancellation is still a necessary criterion. Van Kuppevelt (1996) presents instances in which it is not possible to cancel the implicature all together; for example when the implicature consists of a numeral that is the focus of the utterance. Other authors however believe that the examples presented by van Kuppevelt are not even implicatures in the first place (e.g., Spector, 2013). In summary, we can conclude that because of their uncertain nature scalar implicatures are cancellable. Whether or not this cancellation is successful or not, depends on the specific nature of the scalar implication in question.

Within the current literature, there are two conflicting theories on how scalar implicatures actually work, the Default account and the Context-driven account. The Default account (Chierchia, 2004; Grice, 1991; Grodner, Klein, Carbary, & Tanenhaus, 2010; Horn, 1984; Levinson, 2000)

assumes that the pragmatic interpretation of scalar terms is the default interpretation. The rejection of the stronger term happens automatically. The scalar term will only be interpreted logically when the context indicates that the implicature has to be undone. Therefore, the logical interpretation will be more effortful than the pragmatic default interpretation. The Context-driven account (Carston, 1997; Sperber, Wilson, He, & Ran, 1986) does not believe that the implicature is made by default. According to this theory, it is more cost effectively to start with the logical interpretation that still entails the stronger term. Only when necessary, when the context demands it, a more elaborate interpretation will be constructed. This pragmatic interpretation will not be automatic and will therefore require more energy. From these two conflicting theories, the Default account seems more intuitive to a lot of people. It seems very natural to interpret the word *some* in a way that excludes *all*. However, many experimental findings seem to point in the direction of the Context-driven account. First of all, many studies have shown that children do not have the same pragmatic fluency in interpreting implicatures as adults do (Braine & Romain, 1981; Feeney, Craffton, Duckworth, & Handley, 2004; Guasti et al., 2005; Noveck, 2001; Noveck & Sperber, 2007; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004; Paris, 1973; Pouscoulos, Noveck, Politzer, & Bastide, 2007; Smith, 1980; Sternberg, 1979). This seems to support the idea that the logical interpretation develops first in children and that the pragmatic interpretation only develops later on in childhood. This is in accordance with the Context-driven account. Secondly, Bott and Noveck (2004) have showed that people produce logical interpretations of scalar implicatures faster than pragmatic interpretations. This is also in accordance with the Context-driven account. Faster reaction times in producing logical interpretations point in the

direction that logical interpretations are more automatic and default than pragmatic interpretations. Thirdly, it has also been shown that when participants' cognitive resources are burdened with a secondary task, they appear to become more logical, compared to when they have all their cognitive resources at their disposal (De Neys & Schaeken, 2007). This is perfectly explainable by the Context-driven account. When people have all of their cognitive resources at their disposal, they are able to make the implicature and reason pragmatically. When these resources are burdened however, only the automatic logical responses, which require the least effort, remain and they will not be overruled by the pragmatic ones. Breheny, Katsos, and Williams (2006) conducted a study in which they measured reading times of utterances containing a scalar term. In some utterances it was important to make the scalar implicature to interpret the utterance correctly while in others it was not. They found that reading times of the scalar term were longer when the pragmatic interpretation was optimal. This confirms that making the implicature takes up more time and is not an automated process. A final study conducted by Tomlinson Jr, Bailey, and Bott (2013) used a novel mouse-tracking paradigm that revealed that participants start off with a basic logical interpretation of a scalar term and later on enrich that interpretation towards the scalar implicature. In the experiment, participants had to judge utterances with scalar terms and they had to use their mouse to click either left or right in a, respectively, true or false box. Results showed that the movement towards the 'false' box for the scalar implicature items started off in the direction of the 'true' box and only halfway changed direction towards the 'false' box. It seems the thinking process started off with a logical interpretation and was changed to a pragmatic interpretation later on. This small detour of the mouse track was not observed for the control items.

All these findings point towards a theory that states that the processing costs for a pragmatic interpretation are higher than those for a logical one. However, this theorem has been opposed by Grodner et al. (2010) and Politzer-Ahles and Fiorentino (2013). The results of these studies undermine the conclusion that scalar implicatures are time consuming. Politzer-Ahles and Fiorentino (2013) found in their experiment that reading times for utterances were not longer when the appropriate interpretation of the scalar term was the pragmatic one. In their experiment in which they tracked eye-movement, Grodner et al. (2010) presented participants with two separate pictures and a sentence containing a scalar term. The two pictures both depicted the same scenario, with only a small difference. One of the pictures showed the situation how it would be if the scalar term were interpreted in the logical way, while the other picture showed the scenario with the pragmatic interpretation of the scalar term. The eye-movement data revealed that participants focused their attention on the picture with the pragmatic interpretation of the scalar term, as soon as the scalar term was mentioned, without any delay. These results were challenged by Huang and Snedeker (2009). They could not replicate the findings by Grodner et al.; instead they found that participants waited until they heard the whole utterance before focusing their attention to one picture or the other. The reason for this discrepancy probably lies in the used stimuli, as pointed out by Degen and Tanenhaus (2011). While the study by Grodner et al. only used the scalar terms *all*, *some* and *none*, Huang and Snedeker also used numerical expressions like *exactly two*. Because of these additional terms, the uncertainty of what the meaning of the sentence would be increased. It indicated that the participants needed the whole utterance to interpret the scalar term and again that the pragmatic interpretation

was not an automatic one. A final study that confirmed this hypothesis was conducted by Bott, Bailey, and Grodner (2012). They studied how regular scalar implicatures differed from sentences that contained the term *only some*. They found that correctly rejecting sentences with this expression took less time than interpreting a regular scalar implicature. All these results clearly point into the direction that producing scalar implicatures is a time consuming and costly process and confirm the Context-driven account.

1.1. Overview of the experimental studies

In the present dissertation we will make a more in-depth investigation of scalar implicatures with a focus on developmental studies. There are a few different aspects we will focus on. First of all, we will look at the effects of task difficulty. We will work with several different task paradigms. Secondly, we will look at the effect of response measure on pragmatic reasoning. More specifically, we will look at the effect of responding on a scale, compared to giving binary responses. Thirdly, we will look at reversed scalars, a specific item type that has, to our knowledge, not been investigated in detail in previous implicature work. Reversed scalars are utterances that are logically incorrect. They consist of a statement containing *all*, when the correct term would actually be *some but not all*. An example of this statement is found in Noveck (2001), 'All dogs have spots'. Noveck calls these statements false *all* statements. These items only contain a violation of the logical truth and no pragmatic violation. Therefore, we would expect participants to treat these items as completely false. However, we will find in our studies that these particular items are not just one of the many control items. These three aspects

will be investigated in Chapters 2 through 4. In Chapter 5, we will look at scalar diversity, or how reasoning is dependent on the specific scalar implicatures that are used.

1.2. Chapter 2

Research has shown that children appear to be more logical than adults. Research has also shown however, that children are very susceptible to pragmatic reasoning when certain task features are changed. In the first chapter, we made a comparison between children and adults in two separate experiments. In both experiments we worked with a set of different tasks. Previous research has already shown that when cognitive resources are burdened, mostly by a secondary task, people become less pragmatic. Because we were investigating children, we decided to not make our experiments any more complicated than necessary by adding a secondary task. Instead we opted to work with different tasks that varied in difficulty. We also worked with different response measures. Previous research has shown that both children's and adults' performances are influenced when they have to give their responses on a scale instead of giving binary responses like in most implicature research. It seems that the option of a scale gives a more fine-grained reflection of the thinking process behind scalar implicatures. The main difference between Experiment 1 and Experiment 2 will be that in Experiment 1 we only worked with the scalar response measure, while in Experiment 2, we worked with a within-subjects design and all participants received both the binary response measure and the scalar response measure. Our results were as we expected. First of all, task difficulty made a significant difference on pragmatic reasoning. Especially for children, this effect was large, for

adults it was either small or even not present. It seems that a more difficult task will cause children to be even more logical than adults. Secondly, we found that scalar responses do indeed give a more detailed look into the reasoning process behind implicatures. Like previous research, we found that scalar responses were capable of eliminating the differences between adults and children on scalar implicatures, a difference that remained for the binary response condition. Finally, we also found that reversed scalars, an item type that has not received any attention in previous research, elicit responses that cannot be explained by conventional theories. We found that children treat these particular items very similar to scalar implicatures. Adults however, seem to treat these items like control items when responding binary, but more like scalar implicatures when responding on a scale. This particular finding is important as it sheds lights on how scalar implicatures are formed and on the theories that explain them. Theories predict that scalar implicatures are formed because of a combination of the cooperative principle and the ambiguity between the logical and pragmatic interpretation of the weaker scalar term. If this is the case than there is no reason for the reversed scalar, which only contains the stronger scalar term, to be interpreted in any other way than control items.

1.3. Chapter 3

The goal of the study in Chapter 3 was to further explore reversed scalars. Because we found that the children that were tested in Chapter 2 were still fairly pragmatic, a result that is systematically found in scalar implicatures studies with Dutch-speaking participants, (Banga, Heutinck, Berends, & Hendriks, 2009; De Neys & Schaeken, 2007; Dieussaert, Verk-

erk, Gillard, & Schaeken, 2011; Janssens, Fabry and Schaeken, 2014), we decided to test preschoolers in this experiment. We were interested in how preschoolers would process reversed scalars. In addition, we wanted to compare our results to other literature in which preschoolers were also tested. We again used different tasks, for the same reasons as explained in the previous chapter, and a within-subjects design with two response measures. We did not find an effect of task difficulty. We did not find a main effect of response measure, but we did find an interaction between response measure and item type. It seems that preschoolers are aware of the two-folded interpretation of *some*. This experiment confirmed that children process reversed scalars differently than control items.

1.4. Chapter 4

In Chapter 4 we make a first attempt to explain reversed scalars. Contrary to the pragmatic tolerance hypothesis that has been described in the literature we propose a violation tolerance hypothesis. This hypothesis is broader than the scope of the pragmatic tolerance hypothesis. Not only does the violation tolerance explain scalar implicatures, it also explains reversed scalars. To test this hypothesis, we used a task, which was more abstract than the tasks used in the previous chapters. We predicted that our results could follow two different patterns, which would indicate whether participants reasoned either semantically or pragmatically. These response patterns would also indicate if the pragmatic tolerance hypothesis or the violation tolerance hypothesis would be a better explanation of how people reason. Our results did not support the violation tolerance hypothesis but did not outright reject it either. We

suspect that testing our paradigm on children should bring more clarity on the matter.

1.5. Chapter 5

In the final chapter we investigated scalar diversity. While most studies on scalar implicatures use the flagship example of *all* and *some*, we decided to look at other scales and how they are interpreted. Previous literature has already shown that not all scalar implicatures are processed in the same way. In a first experiment, we tested adults on several different scales. We included a secondary task to burden their cognitive resources to make the participants less pragmatic. We used a between-subjects design with the two response measures. We found an effect of response measure but no effect of load. We did find, like we expected, large differences between the different scales. Notwithstanding our load conditions, participants were still extremely pragmatic on some item types. We therefore conducted a second experiment with preschoolers, whom we expected to be less pragmatic than adults. To simplify the task for preschoolers, we used a smaller set of scales and no load condition. Again, we included a between-subject response measure condition. For the preschoolers, we did not find an effect of response measure, but we did find a clear effect of the different scales. It seems clear that the flagship example of 'some' might not be as representative for scalar implicatures as we all like to pretend.

References

- Baltaxe, C. A. (1977). Pragmatic deficits in the language of autistic adolescents. *Journal of Pediatric Psychology*, 2(4), 176–180.
- Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. *Linguistics in the Netherlands*, 26, 1.
- Borge, S. (2009). Conversational implicatures and cancellability. *Acta Analytica*, 24(2), 149–154.
- Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, 66(1), 123–142.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3), 437–457.
- Braine, M. D., & Romain, B. (1981). Development of comprehension of “or”: Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, 31(1), 46–70.
- Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3), 434–463.
- Carston, R. (1997). Informativeness, relevance and scalar implicature. In R. Carston and S. Uchida (Eds.) *Relevance Theory: Applications and Implications*. (pp. 179–236). Amsterdam: John Benjamins.
- Chierchia, G. (2004). Scalar implicatures, polarity phenomena and the syntax/ pragmatics interface. In A. Belletti (Ed.) *Structures and beyond*, (pp. 39–103). Oxford: Oxford University Press.

- Degen, J., & Tanenhaus, M. K. (2011). Making inferences: the case of scalar implicature processing. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3299–3304). Austin, TX: Cognitive Science Society.
- De Neys, W., & Schaeken, W. (2007). When People Are More Logical Under Cognitive Load. *Experimental Psychology*, 54(2), 128–133.
- De Villiers, J., Stainton, R. J., & Szatmari, P. (2007). Pragmatic abilities in autism spectrum disorder: A case study in philosophy and the empirical. *Midwest Studies in Philosophy*, 31(1), 292–317.
- Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, 64(12), 2352–2367.
- Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 58(2), 121.
- Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts*, (pp. 41–58). New York: Academic Press.
- Grice, H. P. (1991). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42.
- Guasti, M.T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20(5), 667–696.

- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (Ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Huang, Y. T., & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology*, 58(3), 376–415.
- Janssens, L., Fabry, I., Schaeken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychologica Belgica*, 54(4), 374-388.
- Kuppevelt, J. (1996). Inferring from topics. *Linguistics and Philosophy*, 19(4), 393–443.
- Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165–188.
- Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'Scalar Inferences'. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 184-212). Palgrave: Basingstoke.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253–282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12(1), 71–82.
- Paris, S. G. (1973). Comprehension of language connectives and propositional logical relationships. *Journal of Experimental Child Psychology*, 16(2), 278–291.

- Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, 39(4), 607–618.
- Politzer-Ahles, S., & Fiorentino, R. (2013). *The realization of scalar inferences: context sensitivity without processing cost*. In Proceedings of the 26th CUNY Conference on Human Sentence Processing, Columbia, SC.
- Pouscoulous, N., Noveck, I.A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14(4), 347–375.
- Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30(2), 191–205.
- Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, 7(5), 273–294.
- Sperber, D., Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.
- Sternberg, R. J. (1979). Developmental patterns in the encoding and combination of logical connectives. *Journal of Experimental Child Psychology*, 28(3), 469–498.
- Surian, L. (1996). Are children with autism deaf to gricean maxims? *Cognitive Neuropsychiatry*, 1(1), 55–72.
- Tomlinson Jr, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69(1), 18–35.
- Weiner, M. (2006). Are all conversational implicatures cancellable? *Analysis*, 66(290), 127–130.

2

**Children and Pragmatic Implicatures:
A Test of the Pragmatic Tolerance
Hypothesis with Different Tasks**

Katrijn Pipijn
Walter Schaeken

Abstract

The pragmatic tolerance hypothesis (Katsos & Smith, 2010) was originated to explain the difference between children and adults concerning scalar implicatures. The use of a Likert-scale was introduced to test this hypothesis. We conducted two studies in which we compare adults' and children's binary and scalar responses to the same underinformative sentences. We also used three separate tasks to look at the effects of task difficulty on performance. The results show that for children, the more difficult tasks lead to fewer pragmatic responses compared to the easier task, Drawings; for adults this difference is not present. Confirming the study by Katsos and Smith (2010; see also Katsos & Bishop, 2011) both children and adults choose the middle options on the scale more when they are confronted with scalar implicatures and they choose more extreme options for the control sentences. Reversed scalar items however are not interpreted as we expected and this advocates for a more general violation tolerance hypotheses.

2.1. Introduction

Communication is not always as straightforward as one might think. In 1989 Grice published his work on the cooperative principle that was meant to explain how our human interaction could be described. The cooperative principle expects a person to interact in a way that furthers the purpose of a conversation and indicates that a listener expects a speaker to do so. The cooperative principle is the reason why people would use scalar implicatures. A listener can interpret something a speaker says with an implicature. When this happens, the meaning of what that speaker says is not explicitly communicated, but the listener derives it nonetheless from what is said. The utterance is underinformative; more information could have been given but has not. For example when a wife asks her husband whether he'll be home for supper, and the husband answers that he has a meeting that will run late that day, then the husband is using an implicature. His wife will not expect him for dinner. One can assume that she accepts the meeting running late will be the reason, or at least a possible reason, that the husband will not be present at dinner. Nevertheless it is still possible that the husband will appear for dinner, for the implicature is cancellable. It is possible that the husband just meant he would be a little late for dinner; still he would not have lied in his earlier utterance.

One specific form of implicatures is scalar implicatures, which we will focus on in this paper. As the name implies, scalar implicatures consist of words that can be situated on a scale, known as Horn scales (see Horn, 1984). These words range from less informative to more informative, for example a scale containing words like *none*, *some* and *all*. For this particular scale, each word further on the scale contains more elements

of a group. When a speaker uses a specific less informative scalar word in an utterance, it is implicated that the more informative word is not applicable. When a person uses the word *some*, the word *all* would not be appropriate. It is considered a mutual understanding between speaker and recipient that the speaker would have used the more informative word if it were suitable. Nevertheless he deliberately chose to use the less informative word on the scale therefore the more informative is not suitable. For example when the prime minister says 'Some banks are collapsing due to the financial crisis', a citizen can assume that *not all* banks are collapsing due to this crisis, for the expression of *some* implicates *not all*. The citizen presumes that the prime minister would have said 'All banks are collapsing due to the financial crisis' if this were the case. If a few months later the prime minister makes the announcement 'All the banks have collapsed due to the financial crisis', this would not be a withdrawal of his earlier statement. It is a specific characteristic of implicatures is that they are cancellable in only one direction. When a speaker uses the weaker term *some*, it can later easily be corrected to *all*. Yet when a speaker initially uses the stronger term *all*, it is not possible to change it to *some* later on. At least not without admitting one was erroneous the first time. The weaker term *some* entails the stronger term *all* but not vice versa.

When a speaker uses the word *some* in an utterance, there are two different ways to interpret this weak scalar term. The first way is the pragmatic way that was described above. A recipient might produce a scalar implicature and assume that the speaker meant *some and not all* with the statement. Yet another way of interpreting the word *some* is a purely explicit logical interpretation. The explicit meaning of the word *some* is *at least one and possibly all*. Both interpretation of the word are equally

correct and it is the choice of the recipient on how he will interpret it.

We already know from different studies that children and adults interpret scalar implicatures in alternative ways. Noveck (2001) argues that a weak scalar term is understood in its explicit meaning first and will appear first in human development. Only later on the more complex pragmatic meaning will be incorporated. This argument is clearly demonstrated by the results of Noveck's study (2001). He found how children of 7-8 years old and 10-11 years old have acceptance rates of 89% and 85% for sentences that are logically true but pragmatically infelicitous. Adults on the other hand, accept these sentences in only 41% of the cases. This clearly demonstrated how for children the pragmatic meaning of these sentences is not incorporated. While for adults these pragmatic meanings are fully incorporated and are used as the principal criterion to accept or reject sentences. The results also show how these differences between children and adults cannot be explained by the children's limited understanding of words like 'some' and 'all'. For all the different utterances that do not hold a conflict between the logical and the pragmatic meaning, the answering patterns of children and adults are very alike.

The reason for the discrepancy between children and adults is not entirely clear. Various factors seem to contribute to this developmental trend. Noveck explains it by the posterior development of the pragmatic understanding of underinformative sentences. The processing of the pragmatic meaning of underinformative sentences is cognitively much more demanding than the processing of the logical meaning (De Neys & Schaeken, 2007). Children, however, have less available cognitive resources (Gathercole, Pickering, & Ambridge, 2004). Because of this, the pragmatic interpretation is harder to incorporate for children. Another factor that contributes to this is the nature of the task.

Pouscoulous, Noveck, Politzer, & Bastide (2007) reported experiments in which they changed the nature of the task from verbal judgments to action-based judgments. Using small boxes that contained tokens, participants were asked to alter the setting of the tokens to match a statement. They were also allowed to leave a setting as it was. Within the experimental design, children's capability to produce implicatures was much higher than in experiments with verbal judgments. This increased implicature production was found for all ages (4-, 5-, and 7-year-olds as well as adults). Still, the developmental effect was present. These experiments show how the understanding of implicatures can be facilitated in young children by changing task features. Other studies have also shown how changing task features can facilitate children's performance (Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004).

Katsos and Smith (2010) did research on underinformative sentences in children and adults. They proposed the pragmatic tolerance hypothesis to explain for the differences between children and adults as well as differences between adults. The starting point of this hypothesis is that there are different degrees of violations. Several violations can lie within an utterance yet not every violation is equally grave. Participants can and will reject utterances that are a grave violation of the logical truth. Yet they might accept or reject an utterance that only holds a violation of informativeness and thus is an infringement of the cooperative principle. There is no implicit rule on how to deal with pragmatically infelicitous utterances. The threshold of what is and what is not acceptable is individual for each person and is called pragmatic tolerance by Katsos and Smith (2010).

An obvious way to test this hypothesis was adopted by Katsos and Smith (2010; also see Katsos & Bishop, 2011; Katsos, Roqueta, Estevan, & Cummins, 2011). Katsos and Smith (2010) introduced the use of a Likert-

scale to the research on underinformative sentences. A Likert-scale is a bipolar psychometric scale on which a participant can indicate to what extent he agrees or disagrees with a certain statement. Katsos and Bishop (2011) made their participants indicate how much they agreed with utterances containing the words *some* and *all*. Both children and adults clearly rejected utterances that were inherently false and accepted utterances that had an optimal use of the words *some* and *all*. Interestingly, for the underinformative utterances, the answering patterns for children and adults were also very similar, as both groups chose the middle option on a 3-point Likert-scale. This is in strong contrast with Noveck (2001) where the answering patterns for children and adults were much more distinct, notwithstanding the fact that the children in this study were older. Katsos and Smith (2010) explain this finding with the pragmatic tolerance principle. Children appear to be competent pragmatic comprehenders. They do sense the pragmatic violation when underinformative sentences are used. Yet due to their different tolerance levels, they do not experience this violation to be grave enough to be rejected. Therefore, when they are confronted with a binary response option, they will not reject the violation while adults will.

In this paper, we want to explore these results more thoroughly. We will make three hypotheses and investigate these in two separate experiments. First of all, we will vary the task method. Most of the scalar implicatures research relies on people's general knowledge about the world. A prototypical item for example is <Some birds have wings>. Making use of their intrinsic knowledge, people judge this sentence and the scalar term and decide how to interpret it. There are however numerous other ways to test scalar inferences. Pouscoulous et al. (2007) and others taught us that the nature of the task is of great importance.

We expect that when we present different tasks, children will reason more or less pragmatic, depending on the task difficulty. We will apply different methods than those used in most scalar implicature research. We get our inspiration from earlier research on underinformative sentences. For example Newstead (1989, 1995) used Euler Circles and Immediate Inferences in his research. These abstract testing methods should be difficult for children and thus induce more logical reasoning. We also developed more child-friendly materials like in Katsos and Bishop (2011) and Katsos and Smith (2010). The material consisted of drawings that should be easier to interpret and thus induce more pragmatic reasoning in children. However, we expect that all three of these tasks will be easy enough for adults so that they will not lead to differences in pragmatic responses for adults.

Our second hypothesis concerns the differences in pragmatic tolerance between children and adults. In accordance with Katsos and Bishop (2011), we include both children and adults in our research. We aim to replicate the similarities between children and adults when scales are used.

Thirdly, we want to put the pragmatic tolerance principle (Katsos & Smith, 2010) to the test. The pragmatic tolerance hypothesis is based on the general assumption that a scalar term is cancellable in one direction. When a weaker term is used, it can be withdrawn later on and be replaced by the stronger term on the scale without making any violations. This is only possible because of the broad definition of the weaker term, which entails the stronger one. It is not feasible in the other direction. The stronger term does not include the weaker one and replacing the former by the latter is not semantically correct. Therefore, we think that looking at the opposite example of a scalar inference, in which a strong

term is replaced by a weaker one (which we call reversed scalar), is an adequate test for the pragmatic tolerance hypothesis. Katsos and his colleagues did not include this particular example in their studies. Noveck (2001) did include this particular item in his study. Although we are not aware of any statistical tests done on this particular item, we do see that accuracy levels on this item are slightly lower than on control items or compared to adults.

2.2. Experiment 1

In accordance with Katsos and Bishop (2011), we include both children and adults in our study. We used two separate response measure conditions. One group of subjects had to respond binary while the other groups had to respond on a scale. We expected that while adults and children differed in the binary condition, they would respond similarly in the scalar response condition.

We decide to adjust the different types of tasks and make them more conformable. In the experiments carried out by Newstead, participants were shown all the possible Euler Circles patterns together and they then had to indicate which of the scalar sentences were true with respect to them. We changed this task so that participants were always shown one sentence and one pattern at a time. For the Immediate Inferences was set up similar to this, participants could only see one descriptive utterance and one sentence at any given time. As a result of the conformity between the different tasks that we implement, we expect to find correlations between the different tasks, in contrast to Newstead (1995). Furthermore, in order to explore the (pragmatic) tolerance concept more closely, we examined both implicatures and reversed scalars.

2.2.1. Material and methods

Thirty-seven Dutch-speaking children participated in this research (mean age: 10.2, range: 9-11). All children were recruited at youth summer camps. Forty-eight adults participated in the research (mean age: 29, range: 20-58). All adults volunteered to participate in the research. None of the participants had to be excluded due to bad performance.

The children received a pen and paper test. The test started with a cover-up story about a boy named Thomas. Thomas was still learning the Dutch language and the children were to indicate how precise and good his answers were. Children had to indicate their answers on a 5-point Likert-scale, with a happy smiley and a frowning smiley at the ends. For adults the experiment was digitalized. They also had a 5-point Likert-scale to indicate their answers. Three different tests were used. All three tests had the same basic structure. We started each trial with a given situation: a sentence, a figure or a drawing. Then the participants were given a statement about the situation. They were instructed to indicate on the scale how well the statement described the situation given above. For an image of the Likert-scale, see Figure 1.

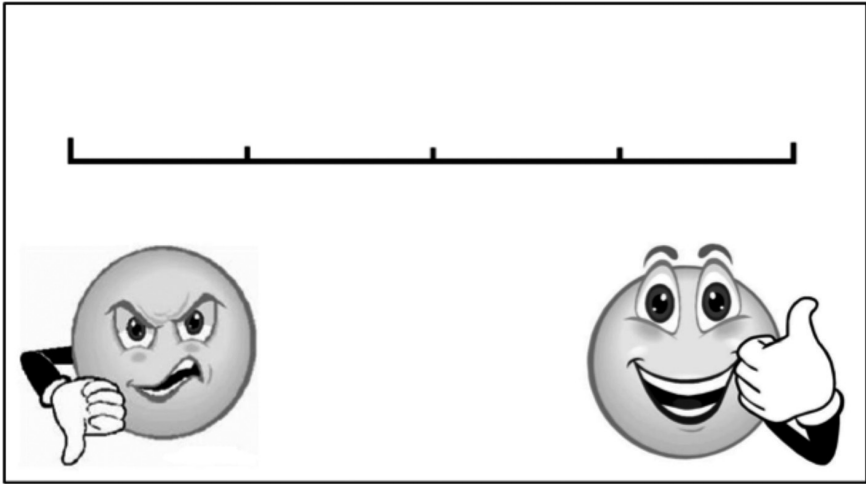


Figure 1. Example of the 5-point Likert-scale.

First was the Immediate Inferences task (II). Each trial started with a descriptive utterance, followed by the statement that the participants had to judge. All the statements were about blocks in different shapes and colors. For example the given utterance was 'all yellow blocks are square' and the questioned statement was 'no yellow blocks are square'. The second task, the Euler Circles task (EC), was very similar, yet the given situation took the shape of Euler Circles. The circles were completely overlapping, partially overlapping or completely disconnected. Each circle represented a group of blocks. Again, the participants received a statement about the circles and they had to judge how precise the statement that described the circles setting was. For an example on this, see Figure 2.

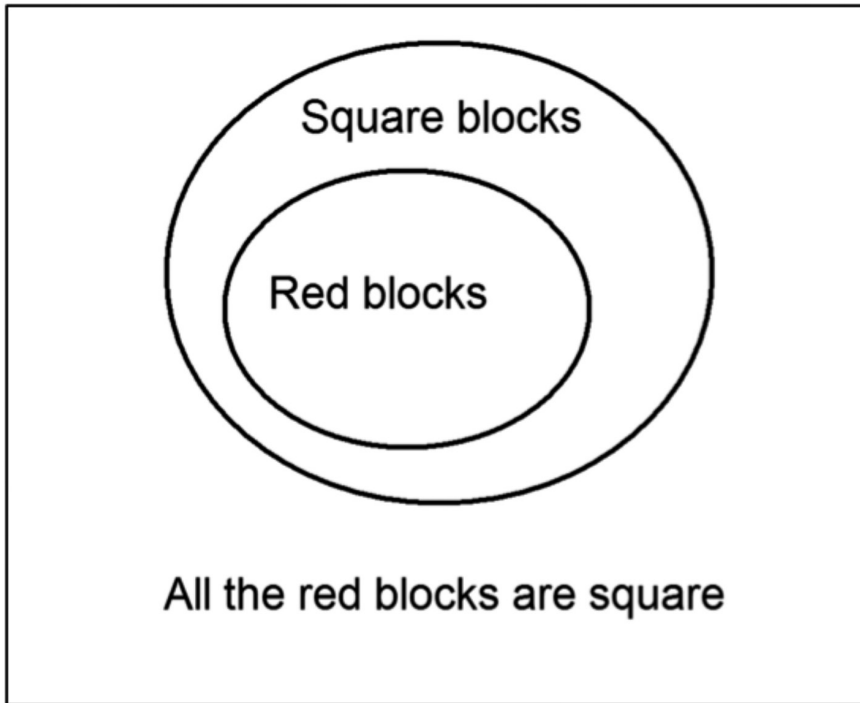


Figure 2. Example of the Euler Circles.

For the third task, Drawings (D), we used our own method that was more adapted to children. For the given situation, the children were now shown a drawing of a real life setting, for example a birthday cake with candles on it. Again the children had to judge a statement about the setting, e.g. 'Some of the candles on the cake are burning'. For an example see Figure 3. Due to the more authentic stimuli, the task became much easier for children.

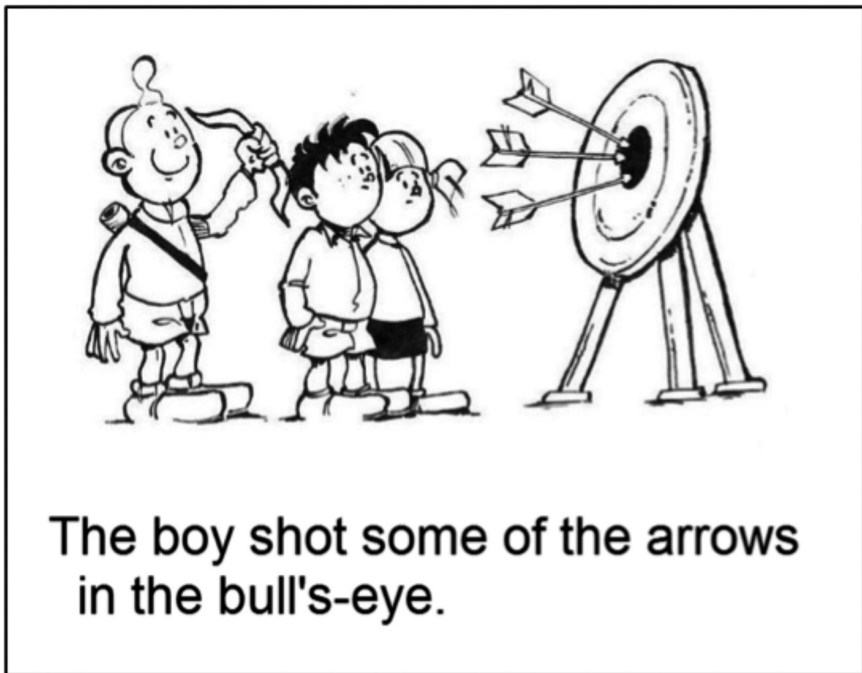


Figure 3. Example of the Drawings.

2.2.2. Results

We inverted all scores of the logically false items. This way, high scores on the control items, for both logically false items and optimal items, indicated competent logical reasoning. We also inverted the implicatures. Because of this, the maximal score of five points is a pragmatic answer and the minimal score of one is a logical answer. We will interpret the pragmatic response as the most optimal and desirable response.

We ran a repeated measures design with one between-subjects variables and two within-subjects variables. The between-subjects variable group consisted of two levels, children and adults. The two within-subjects variables were task and item-type. We had three different

Table 1. Mean scores and standard deviations for tasks and item types.

		Immediate Inferences	Euler circles	Drawings	Total
Children	Control Items	4.19 (.63)	4.19 (.59)	4.54 (.39)	4.31 (.32)
	Implicatures	3.41 (1.26)	2.71 (1.04)	3.66 (1.02)	3.26 (.73)
	Reversed Scalars	3.7 (1.24)	3.05 (1.25)	3.36 (.94)	3.37 (.82)
Adults	Control Items	4.78 (.31)	4.82 (.25)	4.85 (.27)	4.82 (.21)
	Implicatures	2.58 (1.55)	2.96 (1.28)	3.11 (1.46)	2.91 (1.20)
	Reversed Scalars	4.42 (.87)	4.8 (.43)	4.91 (.38)	4.71 (.35)

tasks, II, EC and D. For item types we used control items (C), implicatures (SI) and reversed scalars (RS).

First of all, we found a main effect for task, $F(2, 164) = 9.50, p < .001$. Not all the tasks are equally difficult, as we expected. There is also an interaction between task and age groups, $F(2, 164) = 11.80, p < .001$. Paired t-tests indicated that the II and the EC are more difficult for children than the D task (II-EC: $t(36) = 0.23, p = .98$; II-D: $t(36) = 3.09, p = .004$; EC-D: $t(36) = 3.03, p = .01$). For adults, none of the t-tests showed any significant effects. All mean scores can be found in Table 1.

There is also a main effect for item type, $F(2, 164) = 116.39, p < .001$. Control items, implicatures and reversed scalars were significantly different, (C-SI: $t(84) = 12.51, p < .001$; C-RS: $t(84) = 5.48, p < .001$; SI-RS: $t(84) = 7.63, p < .001$). There is an interaction effect between item type and age groups, $F(2, 164) = 37.36, p < .001$. Both adults and children performed well for the control items. Their average scores are significant-

ly different, $t(83) = 8.82, p < .001$. For the implicatures the average scores did not differ significantly. The manipulation of implicatures versus control items was successful for both children and adults. When confronted with a Likert-scale children and adults answered very alike.

The interaction found between item type and age groups is explained by the reversed scalars. Only for the reversed scalars, adults and children answered differently. We found a significant difference, $t(83) = 10.20, p < .001$. A t-test between implicatures and reversed scalars shows that children interpreted the reversed scalars much like regular implicatures ($t(36) = .897, p = .38$). Adults on the other hand treated these reversed scalars much in the same way as the control items ($t(47) = 1.70, p = .10$).

Finally, we found a three-way interaction between task, item type and group, $F(4, 328) = 4.946, p < .001$.

Next we found correlations between all three tasks (II-EC: $r = .27, p = .01$; II-D: $r = .45, p < .001$; EC-D: $r = .48, p < .001$).

Adults are more consistent in their answers compared to children. We classified participants as consistent when at least two thirds of their answers were similar. We found that 52.1 percent of adults were consistent (either logically, pragmatically or in the middle) but only 24.3 percent of the children were consistent. Within consistent participants, 52 percent of adults is consistently logical, 36 percent pragmatic and 12 percent chooses the middle option. Children and adults show a significantly different pattern of responses. A Fisher's Exact test showed a two-tailed p -value of .01. Thirty-three percent of consistent children chose logical answers, 55 percent was pragmatic and 11 percent took the middle option.

2.2.3. *Discussion*

The results are in line with our expectations. We looked at our three different interests, the comparison between the different methods, the comparison between children and adults and the concept of tolerance.

For the comparison of the three different methods, we found significant correlations between all three tasks. This result is in contrast to Newstead (1995), who could not find a correlation between II and EC. As earlier mentioned, we expect that this correlation became visible due to the fact that we made the different tasks more conformable. Like previously mentioned, we adjusted the way in which the Euler Circles were presented drastically. Our alterations made a direct comparison between the different items less obvious. We believe that this is the reason as to why our results were different from Newstead (1995). Our results show that the different methods do test the same phenomenon. The comparison we made between children and adults was very similar to the one found by Katsos and Bishop (2011). First of all, both adults and children display a significant difference between control items and implicatures. They chose more extreme answers for the control items than for the implicatures. For the control items, children differed significantly from adults. Because of the extreme response patterns for both children and adults, the variances become so small that a significant effect is obtained. For the implicatures, children and adults did not differ significantly. Both groups had average scores close to the middle of the scale. This means that when confronted with a Likert-scale, adults and children reason very alike. The difference between pragmatic adults and logical children disappears when confronted with Likert-scale.

In line with this comparison between adults and children, it does seem necessary to make some remarks. As mentioned earlier, only 25 percent of children were consistent in their answers. It raises the question how many of the children's decisions for implicatures were based on deliberate thinking. Further research on the matter seems necessary. As long as children have these low levels of consistency, it remains difficult to make unambiguous conclusions.

Our final point of interest was the concept of tolerance. Is it the case that children only tolerate pragmatic violations or is it the case that children tolerate less severe violations in general more than adults? We found children to interpret reversed scalars in the same way as implicatures. Adults on the other hand treat them in the way one would expect, the logical way. Neither Katsos and Smith (2010) nor Katsos and Bishop (2011) included this type of items in their experiments. Our findings raise difficulties for the pragmatic tolerance hypothesis by Katsos and Smith (2010). Their hypothesis might have to be interpreted more broadly than the restricted 'pragmatic violations' version. The incorrect interpretations of reversed scalars are not a matter of pragmatic tolerance. It seems that a more general mechanism is at work, a more general violation tolerance hypothesis, a mechanism that is not restricted to pragmatic or logical violations. Children seem to apply the cancellable principle in two directions instead of one.

2.3. Experiment 2

Our second experiment was very similar to the first one. However, we changed our paradigm to a within-subjects design for the response measure variable, instead of a between-subjects design. It seems obvious that the pragmatic tolerance hypothesis should be examined with a within-subjects design in which participants are confronted with a Likert-scale as well as with the two-alternative forced choice paradigm.

We expand the testing method used in Katsos and Bishop (2011). Participants will be confronted with each underinformative sentence twice, once with the option of responding on a Likert-scale or once with a two-alternative forced choice. With this research we expect to replicate Katsos and Bishop's (2011) findings, namely that children do seem to detect a conflict when they are confronted with underinformative sentences. We expect that this conflict detection will be hidden when confronted with a two-alternative forced choice but will become clear when they are confronted with the Likert-scale. Adults on the other hand, we expect to be fairly pragmatic with a two-alternative forced choice. When the scale is introduced however, we expect them to show their feelings of conflict about the inferences by choosing the middle options on the scale. We will use children around the age of eleven, congruent with Noveck (2001). According to this study we expect children of this age to be still much more logical than adults. We will not use the Immediate Inferences task again because we believe that the difference between the Euler Circles task and the Drawings task is large enough that a third task is redundant. We will again look at the reversed scalar and we expect that, congruent to the first experiment that children would treat reversed scalars similar to scalar implicatures while adults will treat them like control items.

2.3.1. Method

2.3.1.1. Subjects

Twenty-two Dutch-speaking children participated in this research (mean age: 11.27; range: 11-13) and 57 adults (mean age 20.51; range 18-57 year). The adults were first year psychology students who participated in exchange for course credit.

2.3.1.2. Procedure

Participants received a pen and paper test. Both children and adults received the same tasks. Adults were also told that the test was originally designed for children, which would explain the childish nature of the task. The test started with a cover-up story about a boy named Thomas. The participants were told that Thomas was new in class and came from a foreign country. They were told he was still learning the Dutch language and the participants were to indicate how precise his answers were. They had to indicate their answers either by indicating right or wrong (two-alternative forced choice), or on a 5-point Likert-scale. The ends of the Likert-scale were illustrated with a happy smiley and a frowning smiley. On the scale, the participants were to indicate how well they thought that the boy's answer was, going from completely wrong to completely right. They were also allowed to use the middle options when the answer was only a little right or wrong or evenly right and wrong.

2.3.1.3. Materials

Two different tasks were used. Both tasks had the same basic structure. We started each trial with a given situation. This situation was presented either by a figure or a drawing. Then the participants were given a statement about the situation. They were instructed to indicate how well the statement described the situation given above.

The first task was the Euler Circles task. Two circles in each figure were completely overlapping, partially overlapping or completely disconnected. Each circle represented a group of blocks, for example 'red blocks', 'square blocks', which was written inside each circle. The participants received a statement about the blocks and had to judge how precise the statement described the circles setting. For an example of this, see Figure 4.

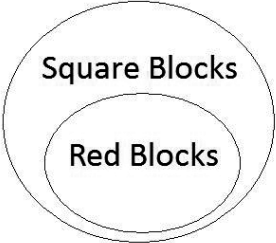
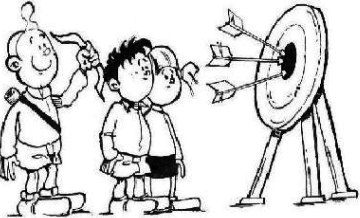

 <p>All the red blocks are square</p>	 <p>The boy shot some of the arrows in the bull's-eye.</p>
<p>The sentence about the blocks is:</p> 	<p>The sentence about the boy is:</p> <p>Right Wrong</p> <p>0 0</p>

Figure 4. Example of Euler Circles, drawings, scalar response option and binary response option.

For the second task, we used a method that was more adapted to children, Drawings. For the given situation, the participants were now shown a drawing of a real life setting, for example a few kids playing with a bow and arrows. Again the children had to judge a statement about the setting, e.g. 'Some arrows are shot in the bull's-eye'. Due to the more authentic stimuli, this second task should, definitely for children, be much easier.

2.3.2. Results

We converted the binary zero and one scores to one and five scores to make them comparable with the scalar responses. The children performed well on both the valid and the invalid control items ($M = 4.59$, $SD = .23$; $M = 1.28$, $SD = .36$). This means that children understand the words *some*, *all* and *none* adequately. For the underinformative items, we found that the average scores were more in the middle of the scale ($M = 2.44$, $SD = .66$). The reversed scalars were higher than we would expect on semantic grounds ($M = 1.97$; $SD = .49$). For more detailed results, see Table 2.

We ran a repeated measures design with three within-subjects variables, and one between-subjects variable: age groups. There was a main effect for age group ($F(1, 77) = 18.77$, $p < .001$). We will explain this effect in light of the different interactions. We found a main effect for response measure ($F(1, 77) = 88.43$, $p < .001$), but no interaction with two groups. We found a significant effect for task ($F(1, 77) = 13.72$, $p < .001$), and also an interaction with age group ($F(1, 77) = 26.49$, $p < .001$). When we look at each task separately, we can see that adults and children perform differently in the Euler Circles task ($t(77) = 4.87$, $p < .001$),

Table 2. Mean ratings and standard errors for all different item types.

			Binary	Scalar
Children	Euler circles	Valid control items	4.38 (.49)	4.24 (.44)
		Invalid control items	1.14 (.35)	1.35 (.63)
		Scalar implicatures	2.82 (1.52)	3.35 (1.05)
		Reversed scalars	1.54 (.67)	2.73 (.88)
	Drawings	Valid control items	4.93 (.24)	4.79 (.33)
		Invalid control items	1.09 (.29)	1.53 (.81)
		Scalar implicatures	1.30 (.57)	2.29 (1.11)
		Reversed scalars	1.06 (.28)	2.55 (.94)
Adults	Euler circles	Valid control items	4.90 (.26)	4.82 (.32)
		Invalid control items	1 (.00)	1.08 (.27)
		Scalar implicatures	1.42 (1.01)	2.32 (.79)
		Reversed scalars	1.05 (.25)	1.77 (.69)
	Drawings	Valid control items	4.93 (.23)	4.88 (.23)
		Invalid control items	1.05 (.23)	1.23 (.33)
		Scalar implicatures	1.23 (.84)	2.61 (.92)
		Reversed scalars	1.05 (.35)	1.71 (.79)

but the same on the Drawings task ($t(77) = 0.94, p = .09$). We also found a difference between the two tasks for children ($t(21) = 4.01, p < .001$), but not for adults ($t(56) = 1.59, p = .12$).

There is a main effect for item ($F(3, 231) = 26.49, p < .001$) and an interaction between item type and age group ($F(3, 231) = 26.49, p < .001$). All four items are significantly different between the two age groups (valid: $t(77) = 6.57, p < .001$; invalid: $t(77) = 2.40, p = .03$; SI: $t(77) = 3.44, p < .001$; RS: $t(77) = 5.89, p < .001$). On Figure 5, we can clearly see that these differences are the largest for scalar implicatures and

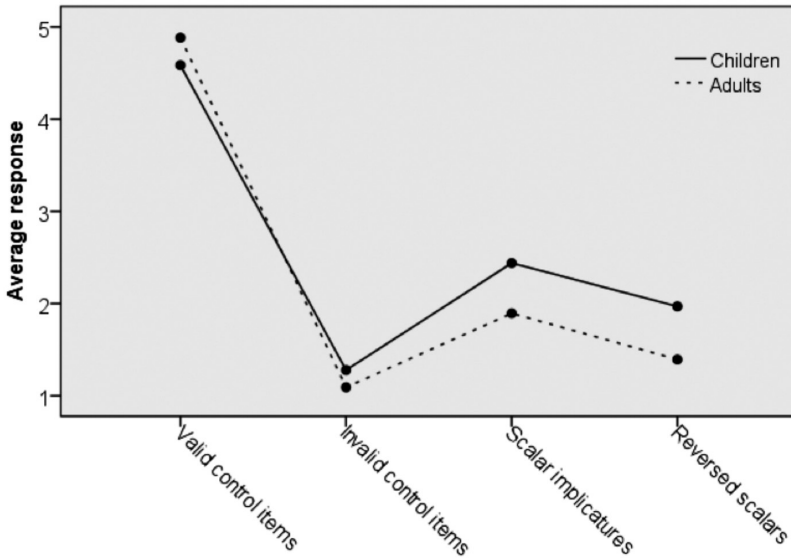


Figure 5. Average responses for the four items types for children and adults.

reversed scalars. We can also see that for scalar implicatures, adults are more pragmatic than children. For the reversed scalars, adults seem to be more logically correct than children, by rejecting the reversed scalars more than children. We also looked at the two groups separately. For children we found, as expected, a significant difference between the two types of control items ($t(21) = 35.91, p < .001$), and between the invalid control items and the scalar implicatures ($t(21) = 7.94, p < .001$). More interesting though, is that also the reversed scalars are different from the invalid control items ($t(21) = 7.08, p < .001$), as well as from the scalar implicatures ($t(21) = 3.12, p < .001$). For adults, we found the same significant differences (valid – invalid: $t(56) = 122.50, p < .001$; invalid - SI: $t(56) = 10.00, p < .001$; invalid – RS: $t(56) = 7.01, p < .001$; SI – RS: $t(56) = 5.88, p < .001$).

There is an interaction between measure method and task ($F(1, 77) = 11.73, p < .001$), but no interaction between response measure, task and age groups. There is a difference between the two tasks when participants have to answer binary ($t(78) = 3.22, p < .001$) but not when a scalar response option is presented ($t(78) = 1.12, p = .27$). When answering binary, the Euler Circles task ($M = 2.20$) seems harder than the Drawings task ($M = 2.07$). However, Figure 6 shows that this significant difference is a rather small difference.

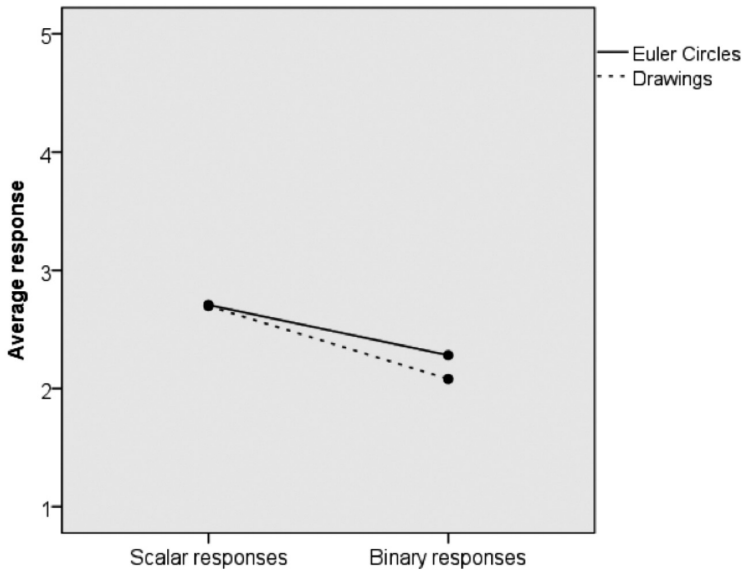


Figure 6. Average responses for the two tasks with the two response measure.

Both the interaction between response measure and item, and the interaction between these two and age groups, were significant ($F(3, 231) = 52.77, p < .001$; $F(3, 231) = 8.22, p < .001$). When looking at the individual t-test, we can see that all pairs of items with different response measures are significant from each other: valid control items: $t(78) = 2.71, p < .001$; invalid control items: $t(78) = 3.93, p < .001$; SI: $t(78) = 8.74, p < .001$; RS: $t(78) = 10.00, p < .001$. Both groups gave extreme answers for the control items for both the response measures. For the scalar implicatures, children gave logical answers when answering both binary and on a scale, while adults only gave logical answers on a scale. When answering binary, adults were extremely pragmatic. To confirm this, we found a significant difference between children and adults for the scalar implicatures when they are answering binary ($t(77) = 3.43, p < .001$), but not when they were answering on a scale ($t(77) = 1.82, p = .73$). For the reversed scalar, both adults and children answered correctly when answering binary but for both groups, performance decreased when responding on a scale. Especially children became very erroneous when responding to reversed scalars on a scale, even more so than adults. Statistically they still significantly differ from each other (scale: $t(77) = 5.04, p < .001$; binary: $t(77) = 3.30, p < .001$). The separate patterns for children and adults can be found in Figures 7 and 8.

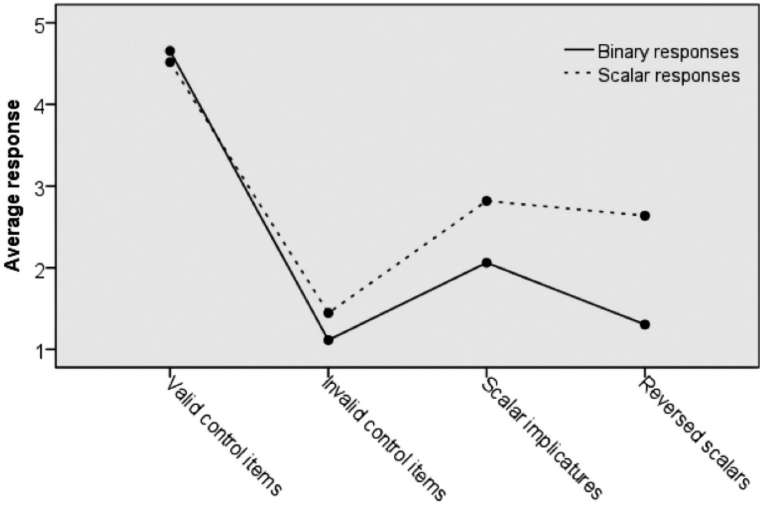


Figure 7. Average responses for children on the four item types for the two response measures.

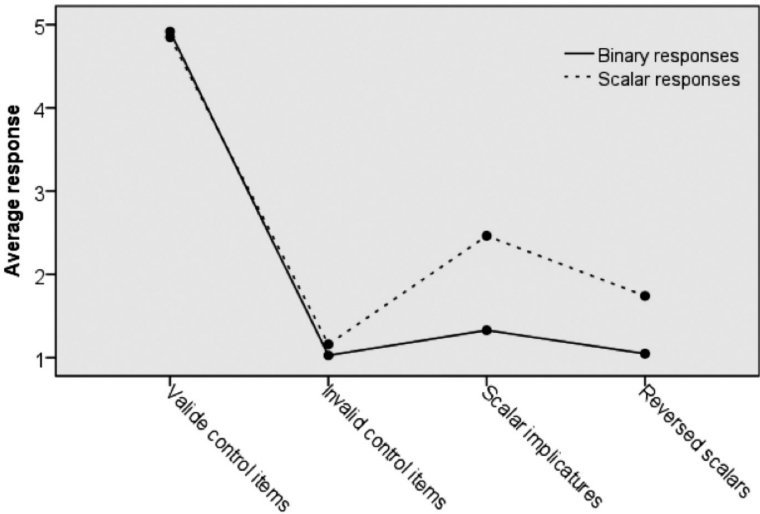


Figure 8. Average responses for adults on the four item types for the two response measures.

A similar pattern is found for the task interaction (Figure 9 and 10). The interaction between task and item type is significant ($F(1, 231) = 35.42, p < .001$), as well as the interaction between task, item type and age groups ($F(3, 231) = 34.38, p < .001$). There seems to be a significant effect of task on most items, but not on the reversed scalars: valid control items: $t(78) = 4.15, p < .001$; invalid control items: $t(78) = 2.60, p = .001$; SI: $t(78) = 2.94, p < .001$; RS: $t(78) = 1.98, p = .05$. When we look at the interaction between tasks and item type for the two groups separately, the effect of task difficulty seems irrelevant for adults. Except for the invalid control items, all pairs of items between the two tasks are not significantly different: valid control items: $t(56) = 1.06, p = .29$; invalid control items: $t(56) = 3.04, p < .001$; SI: $t(56) = .60, p = .55$; RS: $t(56) = .54, p = .59$. Children however, are, especially on the scalar implicatures, influenced by this task difficulty, leading to more logical answers in the Euler Circles task. This task difficulty is then diminished for the reversed scalars (valid control items: $t(21) = 6.92, p < .001$; invalid control items: $t(21) = .72, p = .48$; SI: $t(21) = 6.23, p < .001$; RS: $t(21) = 2.49, p = .02$).

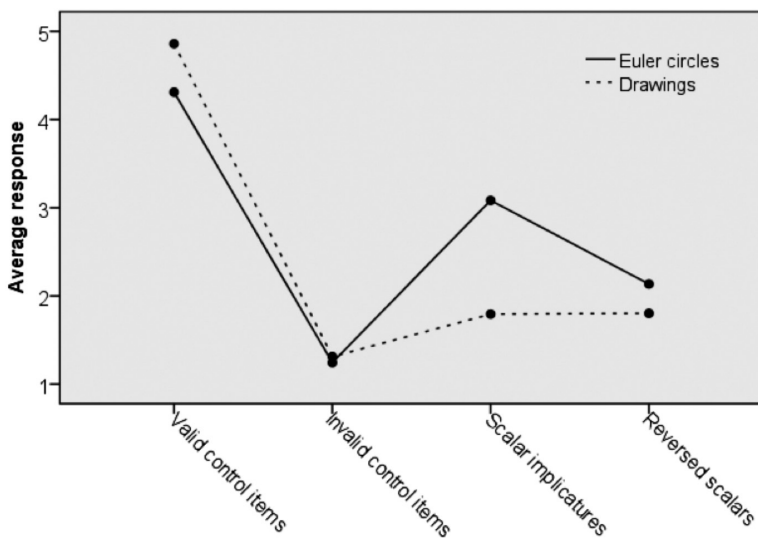


Figure 9. Average response for children on the four item types for the two tasks.

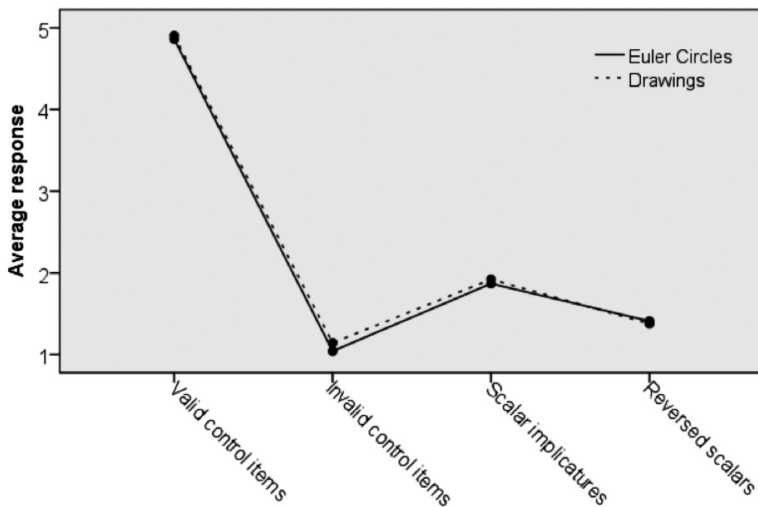


Figure 10. Average responses for adults on the four item types for the two tasks.

There is also a three-way interaction between response measure, task and item ($F(3, 231) = 3.68, p = .001$), but there is no interaction when age groups is added ($F(1, 77) = .84, p = .48$). This three-way interaction can be seen in Figure 11 and 12. We can see that the difference between the two response measures becomes the largest when answering a scalar implicature in the easier Drawings task and smaller when answering this item in the more difficult Euler Circles tasks. No such differences are present for the control items and while there is this difference between the two response measures on reversed scalars, task difficulty does not influence this.

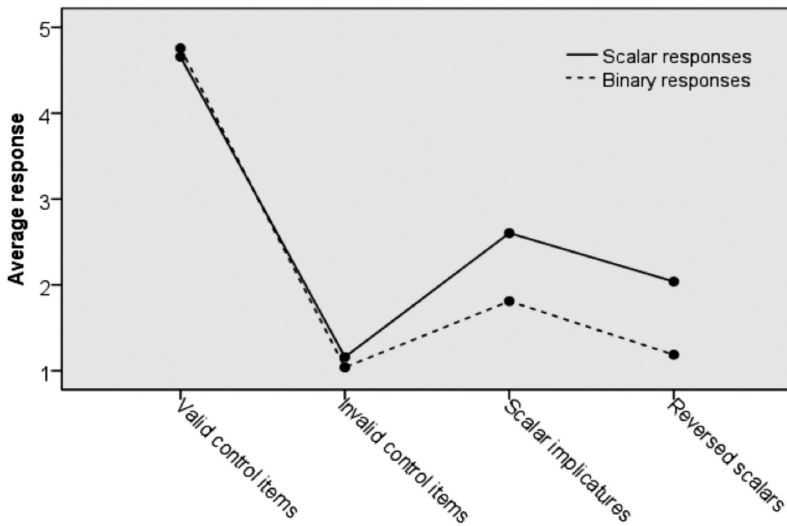


Figure 11. Average responses in the Euler Circles task on the four item types for the two response measures.

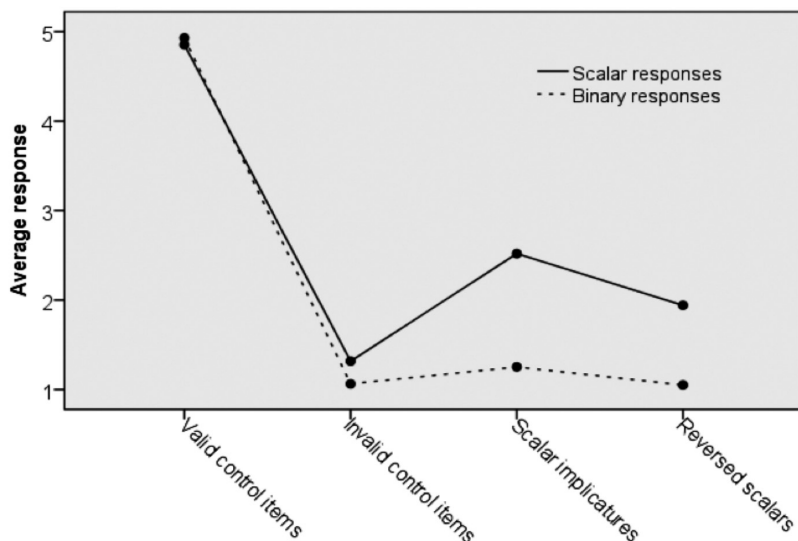


Figure 12. Average responses in the Drawings task on the four item types for the two response measures.

Finally, we look at the distribution of answers across the scales. These are presented in Figure 13. In this figure we put the valid and invalid control items together. We can see that for these control items, both groups answer with extreme answers of the scale. For the scalar implicatures and the reversed scalars, the three middle options are used more frequently. We can see that for children, answers on scalar implicatures are distributed fairly evenly over the five options. Adults however have a slightly skewed distribution in the direction of the pragmatic answers. It seems that both adults and children use the scale correctly to express a feeling of conflict. The distributions for the reversed scalars are even more interesting though. For adults, we see a strong preference for the correct response of 1, some answers for 2 or 3. Children, however, seem

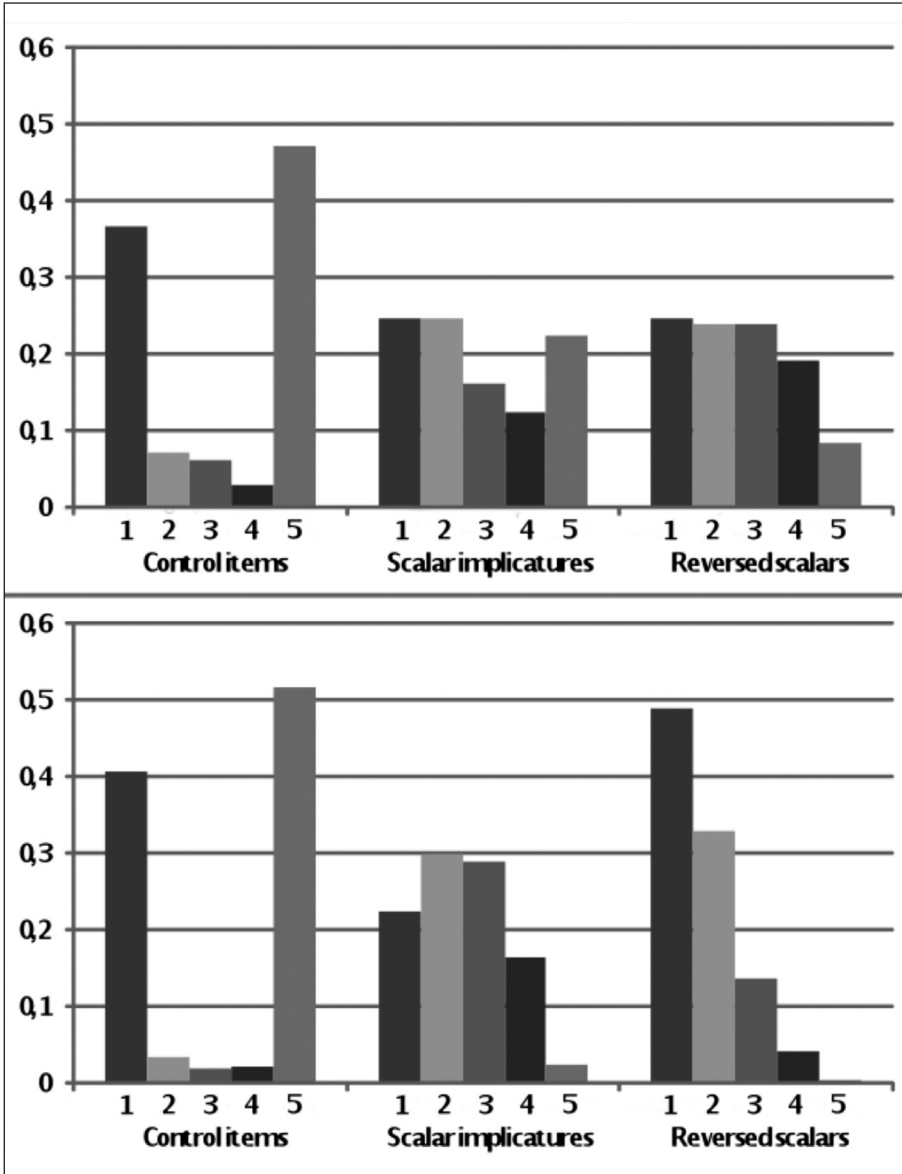


Figure 13. Distributions of scalar responses for the different item types for children (top) and adults (bottom).

to have a similar distribution as for the scalar implicatures, with an even stronger preference for the middle option of the scale. The extreme positive answer on the scale is barely ever picked, but the three middle options take up to 67 percent of all answers. This means that in 75 percent of the cases, children are semantically wrong in interpreting these items.

2.3.3. Discussion

In this second experiment we examined three hypotheses. First of all, we expected that children's performance would depend on the task difficulty and that adults' performance would not be influenced by it. More precisely, we expected the Euler Circles to be more difficult for children than the Drawings task and to lead to fewer pragmatic answers for the underinformative items. We did not expect this difference to be at play with adults. Next, we expected to replicate Katsos and Bishop's (2011) findings, namely that both children and adults answer on the extremities of a scale when confronted with control items but are more doubtful when confronted with underinformative items and a scale. Thirdly, we looked at reversed scalars, for which one would expect both adults and children to reason logically and reject these sentences.

For the first hypothesis, we can find confirmation in the main effect of task for children, and the interaction between task and item type and the interaction between task, item type and age group. For children, the Euler Circles task is clearly more difficult than the Drawings task. Even for the control items this difference is small but significant. For the scalar implicatures, this difference becomes even larger. The more difficult task, the Euler Circles, leads to more logical answers. With the easier task, the Drawings, children become more pragmatic. These pragmatic answers

are still significantly different from the control items though. For adults however, there is no interaction between task and item type. We only found a small significant difference between the two tasks for the invalid control items. However, this difference was very minor and negligible.

Secondly, we found a significant effect of response measure and interaction with item type and an interaction with age groups. The difference between binary answers and scalar answers for the control items is significant. But the difference between the methods becomes much larger for the scalar implicatures. This confirms our hypothesis and replicates Katsos and Bishop (2011). When confronted with a scalar implicature, both children and adults feel that there is a conflict between the pragmatic and the logical interpretation of this scalar implicature. When they can give their response on a scale instead of a binary right or wrong, they have an opportunity to express this feeling of conflict. They tend to choose the three middle options of the scale more often (53% for children and 75% for adults) than when confronted with control items (16% and 7%). Even though we see that children are less inclined to use the middle options of the scale, it does rule out the possibility that children are just unfamiliar with the use of scales. They are adequate in using scales and it is for both groups a deliberate action to choose the middle options for the scalar implicatures and the more extreme options for the control items. This confirms the pragmatic tolerance hypothesis in that both children and adults use the scale to express that they feel the conflict between the logical and the pragmatic interpretation. Children are more logical when interpreting a scalar implicatures binary, compared to adults, but they get much closer to each other when they can answer on a scale.

Most surprising about our results are the reversed scalars. Neither the adults, nor the children, treated them the way one would expect based on semantic rules. Both groups seem to do reasonably well when they can judge the reversed scalars binary. However, when they have to answer on a scale, both groups, children even more than adults, seem to become illogical. About half of the adults and up to 67 percent of children move over to the three middle options on the scale.

2.4. General discussion

Our results show a clear effect of task difficulty. We can hereby confirm what Pouscoulous et al. (2007) and others have claimed. Task features can influence children's pragmatic reasoning on underinformative sentences. We noted earlier that we expect task difficulty to be the determining factor here. De Neys and Schaeken (2007) already showed that cognitive resources are essential for the incorporation of a scalar implicature. When children are confronted with a difficult task, like Euler Circles, the task by itself will take up most of their available resources, leaving little resources for anything else than the basic default logical interpretation. Adults however have a much larger supply of cognitive resources, making the Euler Circles task easier to begin with. Most of them should have plenty of spare cognitive resources to execute a more intensive pragmatic reasoning process. Yet we acknowledge that another factor may be at work as well. The Euler Circles task is believed to rely on logical reasoning skills. It might be possible that the logical interpretation is triggered by the general logical characteristics of the task. In this case, not task difficulty but the logical nature of the task would be the determining factor. The tasks used in this study were also very adapted to usage with

children. More strictly grammatical approaches to the material, instead of the visual child-friendly approach, might lead to different conclusions. More in depth research on the matter seems necessary.

Our results replicate the findings of Katsos and colleagues (Katsos & Bishop 2011; Katsos & Smith, 2010). We did however find a difference with conventional literature on scalar implicatures. The children in this study seem to be much more pragmatic than reports from other studies, especially with the binary responses. One explanation for this is probably the children's ages. Much research on this topic used younger children than the ones used in this study. It is self-evident that the slightly older children used in this study would perform more pragmatically and adult-like. Moreover, the current study was conducted in Dutch. Previous research (Banga, Heutinck, Berends, & Hendriks, 2009; De Neys & Schaeken, 2007; Dieussaert, Verkerk, Gillard, & Schaeken, 2011, Janssens, Fabry and Schaeken, 2014) on underinformative sentences with Dutch speaking children revealed that these children are more pragmatic than their English-speaking (Katsos and Bishop, 2011) or French-speaking (Noveck, 2001) counterparts. Dutch speaking children seem to be more comparable to Spanish speaking children for example. In a study by Katsos et al. (2011), Spanish-speaking children rejected pragmatically false underinformative statements in 87% of the cases. It seems that the Dutch word 'sommige' is not the exact equal of the English word 'some'. This will probably contribute to the high rate of pragmatic answers in Dutch-speaking children.

The finding of the reversed scalars is not only unexpected considering semantic rules, it also opposes the pragmatic tolerance theory. Katsos and Smith (2010) claimed that when a scalar implicatures is presented, people have the option to make a pragmatic violation and treat

the inference logically. However, part of the theory is that this is only possible when a listener has the choice between a pragmatic and a logical interpretation. There lies no such distinction within a reversed scalar. The term *all* is used and there is only one correct interpretation of this word, which is both logically and pragmatically accepted.

The pragmatic tolerance hypothesis might have to be interpreted more broadly than the restricted 'pragmatic violations' version. The incorrect interpretations of reversed scalars are not a matter of pragmatic tolerance. It seems that a more general mechanism is at work, a more general violation tolerance hypothesis, a mechanism that is not restricted to pragmatic or logical violations. When people are confronted with a violation, which could be either logical or pragmatic or both, they will classify it either as correct or incorrect, using their own internal tolerance threshold. This tolerance threshold is individual for each person. When people become older and more and more semantics and pragmatics are incorporated, their personal tolerance threshold gets more fine-tuned. This would lead to a discrepancy with children, who still have a fairly basic knowledge of semantics and pragmatics and thus a fairly crude tolerance threshold. It would be interesting to find out whether and why reversed scalars are treated differently from other logical violations and similar to scalar implicatures. We expect that the gravity of an error will be a main factor in this. The gravity of the errors could be determined by two things: the relative relations between the words of Horn scales or the distance between the presented situation and the word used. If the meanings of two words of a Horn scale are close to each other, then their interchangeable use will be a less grave error than when the words are far apart in meaning. Therefore, higher levels of tolerance can be expected for words closer to each other. We expect that children

are less sensitive to logical errors than adults. Consequently, they will be more tolerant to these errors than adults. It then remains a question whether only the order of the words is important or whether there are also relative distances between words or words and situations that are important.

In conclusion, our study mainly confirms that task features are very important and can influence the rates of pragmatic inferences that are made. The interesting interpretation of reversed scalars puts into question the pragmatic tolerance hypothesis. It is clear to us that the pragmatic tolerance hypothesis and the relationship between binary and scalar answers on underinformative sentences are not as straightforward and that more thorough research on the matter is necessary.

References

- Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. *Linguistics in the Netherlands*, 26, 1.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54, 128-133.
- Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, 64(12), 2352-2367.
- Gathercole, S., Pickering, S., & Ambridge, B. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40, 177-190.
- Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Guasti, M.T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20, 667-696.
- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (Ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Janssens, L., Fabry, I., Schaeken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychologica Belgica*, 54 (4), 374-388.

- Katsos, N., Andrés Roqueta, C., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119, 43–57.
- Katsos, N., & Bishop, D.V.M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, 120, 67–81.
- Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness? In K. Franich, K.M. Iserman, L.L. Keil (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*. Somerville, MA: Cascadilla Press.
- Newstead, S. E. (1989). Interpretational errors in syllogistic reasoning. *Journal of Memory and Language*, 28, 78–91.
- Newstead, S. E. (1995). Gricean implicatures and syllogistic reasoning. *Journal of Memory and Language*, 34, 644–664.
- Noveck, I.A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78, 165–188.
- Papafragou, A., & Musolino J. (2003). Scalar implicatures: experiments at the semantics/pragmatics interface. *Cognition*, 86, 253–282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12, 71–82.
- Pouscoulous, N., Noveck, I., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14, 347–376.

3

**Is it Tolerance or Pragmatic Tolerance?
The Pragmatic Tolerance Hypothesis
in preschoolers**

Katrijn Pipijn
Walter Schaeken

Abstract

Several studies have shown it is much harder for children to interpret scalar implicatures in a pragmatic way than it is for adults. There have also been a lot of studies that have illustrated how both children and adults can be stimulated to reason more pragmatic or more logical. In a previous study we have shown that one way to make children more or less pragmatic is by changing task difficulty. Another way to influence both children and adults is by giving them a scalar response option instead of a binary response option. However, the children that we have studied were still fairly pragmatic. In this article we will investigate a younger group of children: preschoolers. We used two separate tasks that differ in task difficulty to test their pragmatic reasoning skills on scalar implicatures. We expect that more pragmatic responses will be given in the easier task. The children will have to give their responses on a Likert-scale. We expect, congruent to previous experiments, that preschoolers will be aware of the conflict between the logical and the pragmatic interpretation and that they will express this by using the middle options on the scale. We will also look at how preschoolers interpret reversed scalars. We expect them to, even more than the older children from our previous experiments, use the middle of the scale to interpret these items, indicating that they are not regular control items. We did not find an effect of task difficulty in our experiment. We suspect that the tasks that we used were not diverse enough to have any effect. We did find an effect of response measure on pragmatic reasoning. Like we expected, preschoolers seize the opportunity of the scale to indicate the ambiguous nature of scalar implicatures. Finally, also as predicted, we found that reversed scalars were not interpreted correctly. Instead of outright rejecting these

items, children used the middle of the scale for their responses, indicating that these items have a questionable nature.

3.1. Introduction

A growing amount of the research on language and verbal communication focuses on pragmatic inferences. In a normal conversation, a person expresses more than the literal meaning of his words and sentences. Much information is hidden in what a person says and in most cases, a receiver is able to intercept this hidden information and interpret it in the correct way. Some of the most influential work in this area was done by Grice (1991). Grice makes the distinction between what a person literally says and what he actually means, thus, what is implicated. When a receiver is able to recover the hidden added meaning in what a speaker says, it is said that the listener makes an implicature. The receiver is able to make the implicature based on what he knows from the specific context and situation in which the sentence is uttered. Grice believes that this use of implicatures is based on a general cooperative principle. This principle can be divided into several Maxims: *information*, *truth*, *relevance* and *clarity*. Give all the *information* that is required but not more than is required. Give only the *truth*, nothing you believe is false or for which you do not have any evidence. Be *relevant* and *avoid ambiguity*. Conversations work because each person assumes that every other person in the same conversation follows those same Maxims. The listener will always assume a speaker is not violating any Maxims. In some cases, the speaker might use this assumption to get some extra information across by intentionally violating the principle, by using an implicature. We will clarify this with an example:

John: "Are you going to the birthday party?"

Adam: "I have to work".

When Adam gives the answer he does, John can deduce many things from this answer. First of all, that Adam will not make it to the birthday party. The specific reason as to why he will not be able to make it will be because he has to work at the same time as the party. This is much more information than lies in the literal meaning of 'I have to work'. Adam uses an implicature to get his message across and John most likely will understand the implicature and deduce the correct information from what Adam says. Nevertheless it is possible that Adam will show up at the party. Maybe Adam wasn't sure yet whether a meeting would run late or not. He could not confirm or deny his attendance with certainty yet; therefore he chose to use the implicature. A characteristic of implicatures is that they are cancellable. Because the information is not literally stated but only implied, it can easily be retracted later on.

The use of implicatures can be combined with the use of scalar terms, which leads to scalar implicatures. Much of the work done on scalar implicatures was carried out by Horn (1984, 1989). Fundamental to the construction of scalar implicatures are Horn scales. Horn scales are groups of words that can be put next to each other. The meaning of these words can range from weak to strong, with each word on the scale being slightly stronger than the word before it. An example of such a scale is <none, few, some, most, all>. When these words refer to a set of elements, we see that each word further on the scale refers to more elements in that set. While the middle words on this particular scale make a very vague claim as to how many of the elements they refer to, there are also other scales that are much more specific. For example cardinal numbers are also a scale. For the same set of elements, this second type of scale can specify into much more detail how many elements are referred to. We will explain the use of scalar implicatures with an example:

John: "I like *some* movies with Julia Roberts".

In this example John uses the word *some*. The person who hears this sentence will assume from this that John does not like *all* Julia Roberts' movies. Even though the semantic or logical definition of the word *some* is '*at least one, possibly all*', most people will interpret the word as *not all*. This restricted interpretation of the word *some* is the pragmatic interpretation. Both interpretations are correct and it is up to the listener to choose which interpretation he will adopt. The listener will base his decision on the cooperative principle and his knowledge of the context in which the sentence is said. He will assume that the speaker is as informative as possible. Even though a listener knows that the *possibly all* interpretation is correct too, he will assume that if the speaker actually likes *all* the movies with Julia Roberts, he would have used that exact word. Given that the speaker uses the weaker word to express his thoughts, the stronger word must be inaccurate. Again this implicature is cancellable. It is possible that John has not seen them all, but if he did, he actually would have liked all of the Julia Roberts' movies.

The development of pragmatic reasoning has been studied in various studies. Developmental research shows that the representation of weak scalar terms is initially logical and only later on more pragmatic interpretations get incorporated. Smith (1980) conducted a study in which he tested the pragmatic reasoning skills in 4- to 7-year-old children. The children were clearly competed in their general understanding on scalable terms. However, they did not seem to mind treating the word *some* as compatible with *all* in questions like 'Do some birds have wings?'. A study by Noveck (2001) was very similar to this study. In a comparison between 8- and 10-year-old French-speaking children and adults (Exper-

iment 3), he found that adults reasoned much more pragmatically than children. While adults rejected underinformative sentences like 'Some giraffes have a long neck' in 59 percent of the cases, 8- and 10-year-old children only rejected these sentences in, respectively, 11 and 15 percent of the cases. Guasti et al. (2005) replicated the previous study with Italian speaking participants (Experiment 1). The results were almost identical. Children rejected infelicitous sentences in 13 percent of the cases while adults rejected them in 50 percent of the cases.

Papafragou and Musolino (2003) ran an experiment (Experiment 1) in which they compared 5-year-old Greek speaking children to adults. They found rejection rates of 12.5 percent for children and 92.5 percent for adults. The difference between children and adults is even larger in this study, compared to the study by Noveck. We assume that this is due to the lower age of the children in this study compared to the 8- and 10-year olds in Noveck. Previous studies clearly show there is a developmental trend. Children become better at pragmatic reasoning as they get older.

Aside from increasing age, there are other ways to make children's performances closer to adults' performances. Several studies have unraveled ways to make children more sensitive to violations of informativeness. In the second experiment carried out by Papafragou and Musolino (2003), children first ran through a training phase in which they were trained to better detect infelicitous items. After this training phase, rejection of scalar implicatures went up from 12.5 percent to 52.5 percent. The enhancement effect was confirmed by Guasti et al. (2005). Following up on their replication of Noveck (2001), they carried out a second almost identical experiment, in which they included an explicit training phase, comparable to the one in Papafragou and Musolino. The rate at

which children rejected the scalar implicatures went up to 52 percent. However, this rejection rate dropped again to 22 percent one week after the initial testing and training. In a fourth experiment Guasti et al. changed the context in which the scalar implicatures were presented. In this experiment they used the same task, but this time, the stories behind the scalar implicatures were acted out further. This led to an elevation in rejection rates, for both children (73 percent) and adults (83 percent). This context enrichment leads to a reduction of the difference between children and adults. Foppolo, Guasti, and Chierchia (2012) also tested multiple age groups of children. They also found a clear developmental trend. While 4- and 5-year-old children only rejected underinformative items in 42 percent of the cases, 6- and 7-year olds and adults rejected them in around 80 percent of the cases. They were, however, also able to increase the rejection rates of the youngest children up to 72.5 percent by changing the critical word 'some' to 'some of'.

All these tasks clearly show that both the age of children and the way a task is presented lead to large differences in how scalar implicatures are interpreted. Another way to improve children's behavior is by implementing different ways in which they can give their answer. A well-known example of this is by letting children give action-based responses (Pouscoulous, Noveck, Politzer, & Bastide 2007). In their experiment they were able to increase scalar implicature production by giving 4-, 5-, and 7-year olds as well as adults an action-based task. In an elegant simple task they requested that *all/some/no* boxes presented in front of the participants contain a token. When in the pre-trial set-up of the boxes all the boxes already contained a token, and participants were asked that some of the boxes had tokens, they had two options. When they reasoned pragmatically, they would take away at least one token. When

they reasoned logically, they could leave the set-up as it was. This task increased implicature production across all ages. In 68 percent of the cases, 4-year-old children would change the tokens and reason pragmatically, as opposed to 9 percent in a classic truth value judgment task.

Another way to alter the way participants responded was implemented by Katsos and colleagues (Katsos & Bishop, 2011; Katsos, Roqueta, Estevan, & Cummins, 2011; Katsos & Smith, 2010). Instead of answering binary like in a classic truth value judgment task, participants were now asked to answer on a Likert-scale. By doing this, Katsos and Bishop (2011) were able to increase the pragmatic responses in 5- and 6- year-old children from 26 percent (Experiment 1) up to 100 percent (Experiment 2). From this 100 percent, 89 percent was placed on the middle of the 3-point scale. This shows that children do not necessarily reject these pragmatically incorrect sentences, but that they are in fact sensitive to the ambiguous nature of the items. These results were also replicated with 6- and 7- year-old children (Katsos & Smith, 2010). Katsos and Bishop also made a comparison to adults and for adults they found that they chose the first, middle and third option on the scale for the incorrect control items, scalar implicatures and correct control items respectively. They also found that there was no longer a significant effect between the children and adults on this task. Katsos and Smith (2010) explained this effect with the Pragmatic Tolerance hypothesis. They believe that even young children are in fact competent pragmatic comprehenders, but they have different tolerance thresholds. Based on this threshold, children and adults make decisions about which pragmatic violations are acceptable and which are not. There is a difference between children and adults for the strictness of this threshold, with that of children being much more loose and tolerable. Some particular pragmatic violations

will be more acceptable to children compared to adults, while others will not. However, when participants have to answer binary, this tolerance towards pragmatic violations will be hidden. Katsos and Smith believe that it is not necessarily the children's pragmatic reasoning skills that improve with age, but more that their individual pragmatic violations threshold further develops towards that of adults.

In a previous study we replicated these findings from Katsos and colleagues. In the first experiment of our study (Pipijn & Schaeken, 2012), we conducted a study between Dutch-speaking 9- to 11-year-old children and adults, in which they gave participants a 5-point scale to rate the sentences. However, they changed the types of tasks in which the scalar implicatures were presented. Parallel to Newstead (1989, 1995), two tasks were included in the experiment: Immediate Inferences and Euler Circles. Another task, Drawings, was also included in the experiment. These three tasks vary in their difficulty level. The authors expected, especially for children, that these differences in difficulty would lead to differences in the performance on scalar implicatures. More precisely, they expected the more difficult tasks, Immediate Inferences and Euler Circles, to lead to fewer pragmatic answers. The results were as expected, the easiest task, Drawings, led to the most pragmatic answers for children. For adults, there was no effect of task difficulty. The comparison between adults and children was the same as in Katsos and Bishop (2011), children and adults perform similarly on the scalar implicatures when they could answer on a scale.

In the second experiment of our study (Pipijn & Schaeken, 2012), we conducted a very similar experiment, but this time a within subjects comparison was made. Again, adults were compared to 11- to 13-year-old children. However, this time all subjects had to respond to

each item twice, once in a two-alternative forced choice task, and once on a 5-point Likert-scale. The Euler Circles task and the Drawings task were used to look at the effect of task difficulty. When answering binary in the two-alternative forced choice task, children reasoned more logically about scalar implicatures than adults. However, when these same subjects had to rate these same items on a 5-point scale, the difference between adults and children disappeared. Adults in this experiment were not influenced by the task difficulty. Children however, became more logical in the more difficult task, Euler Circles, and more pragmatic in the easier task, Drawings. All these results are in line with one could expect from the literature.

One aspect of these two studies was unexpected though. In both studies, the opposite of a scalar implicatures was included, which is referred to as a reversed scalar. Reversed scalars occur when a *some*-situation is presented and a participant is asked whether an *all*-sentence would be a good description of this situation. For example when someone is asked whether "All teachers have glasses", this sentence is obvious false as only 'some teachers have glasses'. Both the semantic and pragmatic definition of the word *all* state that it entails a whole quantity, each and every one. When one or more elements are missing, the word *all* does not apply anymore. Therefore one would assume that these types of sentences are always rejected in both the experiments described above and would not be different from the other used control items. However, they were not. In our previous study, the judgment of reversed scalars by children was significantly different from the judgment of the incorrect control items. In the first experiment, there was no significant difference between the reversed scalars and the actual scalar implicatures. For adults, the reversed scalars were not significantly dif-

ferent from the incorrect control items. In the second experiment, both children and adults rated the reversed scalars somewhere between the scalar implicatures and the incorrect control items. However, this only occurred when participants had to answer on a scale. When participants had to answer binary, judgments dropped considerably for both adults and children. One could argue that this is due to a limited understanding of the terms *some* and *all*, but results on the control items clearly show that both adults and children are adequate comprehenders of these scalar terms and of the task in general.

In this paper, we want to further explore these findings. To do this, we will look at the performance of preschoolers in these tasks. We will test 5-year-old children in our study and we will use the same tasks as used in the two previous experiments: Euler Circles and Drawings. Again, we will make a within-subjects comparison with the two response methods: binary and scales. The main motivation for testing children of this age is the generally high level of pragmatic reasoning in Dutch-speaking children. The previous two experiments clearly showed that Dutch-speaking children are more pragmatic than their French, Greek or Italian counterparts. Subtle language differences seem to lead to different interpretations of the words *some* (English), *sommige* (Dutch), *certain* (French), *qualche* (Italian) and *Merika* (Greek). We believe it would be interesting to make a comparison between the previous two experiments in which adults and older children were tested, and the preschoolers in the current experiment. Secondly, we also believe it would also be interesting to make a comparison with the 5-year-old children that were tested in Katsos and Bishop (2011). We want to see if we can replicate their findings, and again, we want to make the within-subjects comparison that those authors did

not. We think it will also be interesting to see how Dutch-speaking children respond compared to children in other languages.

We expect to replicate the findings from Katsos and Bishop (2011) and our own previous experiments, namely that children, even as young as 5 years old, do seem to have the ability to reason pragmatically when the task is adjusted. We expect there will be a significant difference between their binary responses and scalar responses, with the scalar responses being more pragmatic than the binary. We also expect the responses to be less pragmatic overall than those of the older children in our previous studies. We also expect to find a task difficulty effect, with the more difficult task, Euler Circles, leading to fewer pragmatic responses than the easier task, Drawings. Finally, we expect these young children to treat reversed scalars differently from other control items. We expect this effect to be even more robust than it was for older children.

3.2. Method

3.2.1. Participants

Thirty-four Dutch-speaking preschoolers participated in the experiment, all assembled from the same elementary school. The children were between the ages of five and seven years old ($M = 5.38$, $SD = .55$). Eighteen girls and sixteen boys participated. We had to exclude five children from the analysis due to bad performance on the control items. All the children that were included in the analysis had at least 75 percent of the control items correct.

3.2.2. Procedure

The experiments were done individual. The children were told a cover story with a doll called 'Professor Bamboozle'. They were said that the professor was going to say some sentences. The professor would sometimes try to fool them by saying sentences that were not right. The children were to indicate if the professor said something right or wrong. Then they could reward him by giving him candy. They could choose to give him 0 to 4 pieces of candy, depending on how right or wrong they thought the sentence was. Each child was presented each item twice, once when they had to reward him with candy, and once when they had to say whether the professor said something right or wrong. There were two separate tasks, the Euler Circles task and the Drawings. The Euler Circles task was always presented first, the Drawings second. This was done deliberately because of the limited attention span of 5-year olds and the difficulty levels of the two tasks. The number of items was limited to make sure the task was not too long for the children.

3.2.3. Materials

In the Euler Circles task, the children were presented with two large circles in which several Attribute Logic Blocks were put down. The two circles were either completely separated, one completely within the other, or partly overlapping. Professor Bamboozle would say a sentence about the blocks laid out in front on the children. Children then had to say how well that sentence described the situation in front of them. To make the task easier and more appropriate for 5-year-old children we used actual block instead of the more abstract version of the task with words like

'Red blocks', like it was done in previous experiments (Pipijn & Schaeken, 2012). With this alteration, the use of the circles became redundant. However, we kept the circles in the experiment to make the task as similar as possible with those previous studies to make a comparison. There were six control items; these were either inherently correct or incorrect. We also presented two scalar implicatures and two reversed scalars. An example of a scalar implicature can be seen in Figure 1. The correct sentence for this example would be 'All round blocks are red'. The use of the word *some* is underinformative in this example.

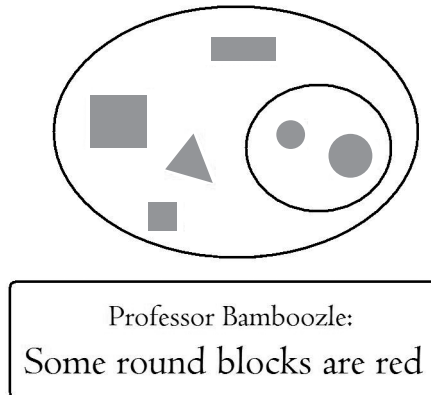


Figure 1. Example of a scalar implicature in the Euler Circles task.

An example of a reversed scalar can be seen in Figure 2. In this example, the correct sentence would be 'Some round blocks are red'. The use of the word *all* is wrong, because the meaning of word is broader than the situation actually is.

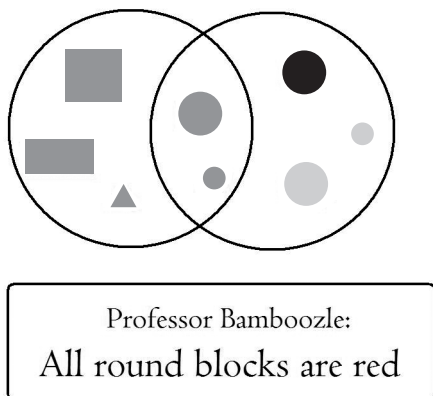


Figure 2. Example of a reversed scalar in the Euler Circles task.

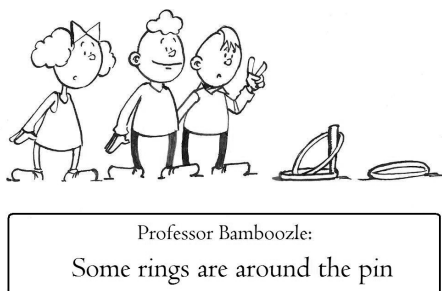


Figure 3. Example of a valid control items in the Drawings task.

For the Drawings task we used child-friendly drawings in which a small situation is outlined. Then Professor Bamboozle asks a question about the drawing. Nine control items were included, which were clearly right of clearly wrong. Figure 3 is an example of such a correct control item. We included three scalar implicatures and three reversed scalars.

We had a total of 25 items and each items had to be rated twice, once when the children had to answer with right of wrong (binary response) and once when they could give the more nuanced answer (scalar response).

3.3. Results

The binary zero and one scores were converted to one and five scores to make them comparable with the scalar responses. The children performed well on both the valid and the invalid control items ($M = 4.74$, $SD = .29$; $M = 1.24$, $SD = .25$). This means that the children understand the words *some*, *all* and *none* adequately. For the underinformative items, we found that the average scores were more in the middle but slightly leaning towards logical answers ($M = 3.73$, $SD = .89$). The reversed scalars were also rated relatively high, which is unexpected based on semantic grounds ($M = 2.83$; $SD = .91$). For more detailed results, see Table 1.

Table 1. Mean ratings and standard errors for all different item types.

	Scalar responses		Binary responses	
	Euler circles	Drawings	Euler circles	Drawings
Valid control Items	4.51 (.59)	4.85 (.20)	4.64 (.51)	4.97 (.15)
Invalid control Items	1.17 (.38)	1.48 (.48)	1.00 (.00)	1.31 (.54)
Scalar implicatures	3.76 (1.20)	3.59 (1.24)	3.97 (1.38)	3.62 (1.53)
Reversed scalars	3.19 (1.15)	2.84 (.76)	3.00 (1.51)	2.29 (1.49)

We ran a repeated measures design with three within-subjects factors, namely response measure (binary vs. scalar), task (Euler Circles vs. Drawings) and item type (valid control items, invalid control items, scalar implicatures, reversed scalars). There were no main effects for response measure or task. There was however a main effect for item ($F(3, 84) = 175.85, p < .001$). We compared the different items individually and they were all significantly different from each other (see Table 2). As expected, the valid and the invalid control items were significantly different from each other. More interesting is that the reversed scalars are significantly different from the invalid control items. The scalar implicatures were significantly different from all other item types.

Table 2. T-test between all different item types and the significance levels.

	t (28)	p-value
Valid - Invalid	45.51	< .001
Valid - SI	6.07	< .001
Valid - RS	11.36	< .001
Invalid - SI	14.22	< .001
Invalid - RS	10.52	< .001
SI - RS	4.82	< .001

There was no significant interaction between response measure and task. There was a significant interaction between response measure and item type ($F(3,84) = 7.54, p < .001$). We looked at the t -test between the two measure methods for each individual item. There was a significant difference between the two response measures for the control items and for the reversed scalars (valid control items: $t(28) = 4.24, p < .001$; invalid control items: $t(28) = 2.76, p < .001$; reversed scalars: $t(28)$

= 2.50, $p = .02$). There was no significant difference between the two response measures on the scalar implicature items ($t(28) = 1.00, p = .33$).

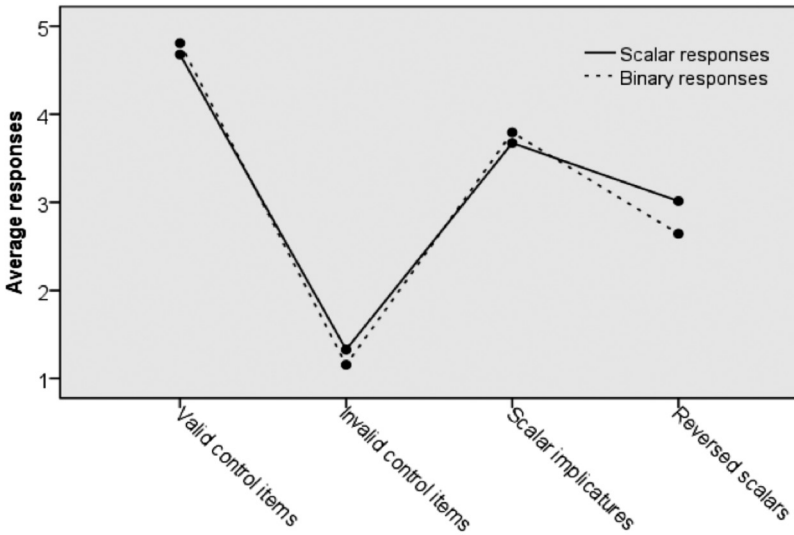


Figure 4. Average responses on all item types for the two response measures.

The interaction between task and item type was also significant ($F(3,84) = 19.89, p < .001$). When looked at the different t -tests, it seems that task makes a difference on the control items and a marginal difference on the reversed scalars, but not on the scalar implicatures (valid control items: $t(28) = 3.43, p < .001$; invalid control items: $t(28) = 3.69, p < .001$; scalar implicatures: $t(28) = 0.76, p = .46$; reversed scalars: $t(28) = 2.02, p = .05$).

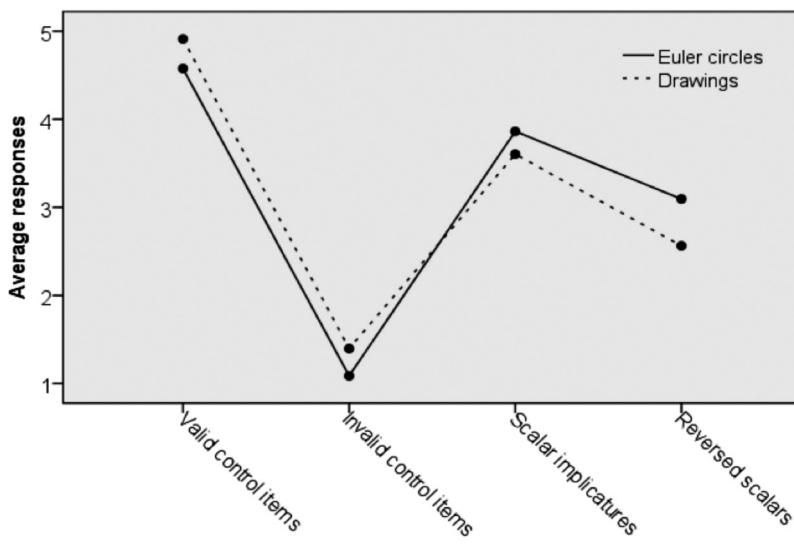


Figure 5. Average responses on all item types for the two tasks.

Finally, we did not find a significant three-way interaction between response measure, task and item type ($F(3, 84) = 1.01, p < .39$). We also looked at the distribution of responses across the scale, for the scalar response measure condition. For the control items, children answer mainly on the two extreme ends of the scale. We see an excess number of answers on five compared to one, but this is due to an unequal number of valid and invalid control items. For the scalar implicatures, we see that in 63 percent of the cases, children still choose the extreme values. Most interesting is the distribution for the reversed scalars. Purely semantically, we would expect most of the answers to be with one, which is the correct answer for these items. However, only 20 percent of children choose the correct item. The middle answer on the scale is chosen the most. This clearly indicates that for 5-year-old children, reversed scalar items are not necessarily wrong.

Table 3. Distribution of responses for the different item types.

	Controle items	Scalar implicatures	Reversed scalars
1	.28	.17	.20
2	.04	.07	.09
3	.05	.15	.38
4	.06	.14	.19
5	.57	.46	.14

3.4. Discussion

We had three main hypotheses for this study. First of all, we expected to find an effect of task difficulty. We expected more difficult tasks to lead to fewer pragmatic responses on scalar implicatures. Secondly, we expected an effect of response measure. We expected the children in our experiment to become more pragmatic when they had to answer on a scale instead of binary. We expected these results to be similar to the results from Katsos and Bishop (2011). We also expect the 5-year-old children tested in this experiment to be less pragmatic than older children tested in other experiments. Finally, we expected children to treat reversed scalars differently from other control items.

We did not find a main effect of task, but we did find an interaction of task with item type. T-tests revealed that this interaction was due to differences between the two tasks on the control items and on the reversed scalars. However, for the scalar implicatures, where we expected to find a difference, there was no difference between the two tasks. On

Figure 5 we saw that the easier task, Drawings, led to more pragmatic responses but this effect was not significant. We could hypothesize that the two tasks that we used in the experiment were maybe too difficult for 5-year-old children. However, we only used the children that performed well on the control items. It is possible that by excluding the children that performed poorly on the control items, we excluded those children for whom task difficulty would matter. Therefore we ran an analysis with the lower half of the participants but this did not change the results. Maybe it is possible that the tasks were not too difficult for the children, but still difficult enough that the burdening was at maximum and no extra resources were available to make the difference between the two tasks. But again this is very unlikely. With average scores between 3.5 and 4, still a fair number of the responses given were pragmatic. It seems that the children did have enough resources available to make some pragmatic responses. It is not clear to us why then we did not find an effect of task on the scalar implicatures while we did find it with older children. We would expect the effect of task difficulty to become larger for younger children for whom the tasks are more difficult than for the older children. The only reasonable argument would be that the two tasks were equal on the difficulty level. Even though we did find significant effects of task on the control items, the direction of these effects was opposite for the valid compared to the invalid control items. A *t*-test, in which the two types of control items are combined, does not show an effect of task on control items any more. The reason as to why we did not find a task effect while it was present in previous studies might be the alterations in the Euler Circles task. We altered the task by changing the abstract descriptions of blocks by actual blocks. This made the task easier for the 5-year-old children, but it is very likely that this also made the difference

in task difficulty between the Euler Circles task and the Drawings task completely disappear. The way we designed the task, the actual circles became redundant and the decision could be made solely on the blocks. The circles might even have been an extra aid that made logical reasoning less necessary and that made the task easier for the children. When there is no difficulty difference any more between the two tasks, it is only natural that there is no effect of task either.

The results that we found concerning response measure were very similar to the results found in Katsos and Bishop (2011). Children outright rejected 17 percent of the scalar implicatures when answering on the scale. We found an interaction effect of response measure and item type. However, a more in-depth analysis of this interaction revealed that there was no response measure effect for the scalar implicatures. When answering binary, children were still fairly logical with 69 percent of all answers being logical. If we include every answer from 1 to 3 on the scale as an indication of a pragmatic answer, logical answers are reduced to 24 percent. Even though there is no effect of measure on scalar implicatures, we do interpret these numbers in the sense that response measure does in fact have its impact. When confronted with a scale, children will express their conflicting feelings about the scalar implicatures by using the full range of the scale. One could question whether this is in fact a deliberate strategy instead of the expression of increased uncertainty caused by the scale. Perhaps a 5-point scale is too difficult for these young children. Katsos and Smith (2010) also used a 5-point scale but they used this scale with older children. In Katsos and Bishop (2011) they also used 5-year-old children, but then they used a 3-point scale. It is possible that the combination of 5-year-old children and a 5-point scale was not an optimal decision. However, the distribution of the responses

for the control items clearly shows that even 5-year-old children are proficient in using this scale to answer and the use of the middle options on the scale does seem like a deliberate strategy.

This deliberate strategy is even further emphasized by the reversed scalars. Even though these particular items are semantically speaking control items, children judge these items significantly different from the other control items. Like we expected from previous studies, the responses on the reversed scalars lie closer to the scalar implicatures than to the control items. In this study, 80 percent of the answers are actually incorrect, with a clear preference for the middle option on the scale. These results put into question what we assume about scalar implicatures, namely that one can only make the implicature in one direction and not the other. Children in this study and in the previous one, clearly use the words *some* and *all* interchangeable in two directions instead of one. It also puts into question the pragmatic tolerance hypotheses by Katsos and Smith (2010). Partly our results confirm their speculation; children do seem to have an early understanding of pragmatics. However, it does not seem to be an individual threshold for pragmatic violations that develops over age. Instead, we believe it is more a threshold for violations in general that develops over age. It seems that especially young children have a very loose opinion about how much one should follow semantic rules, about what is acceptable and what is not. We propose a more general violation tolerance principle.

References

- Foppolo, F., Guasti, M.T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language Learning and Development*, 8(4), 365–394.
- Grice, H. P. (1991). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Guasti, M.T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20(5), 667–696.
- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (Ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Horn, L. (1989). *A natural history of negation*. Chicago: University of Chicago Press.
- Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.
- Katsos, N., Roqueta, C. A., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119(1), 43–57.
- Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness. In K. Franich, K.M. Iserman, L.L. Keil (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*. Somerville, MA: Cascadilla Press.

- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165–188.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253–282.
- Pijp, K. & Schaeken, W. (2012). Children and pragmatic implicatures: A test of the pragmatic tolerance hypothesis with different tasks.
- Pouscoulous, N., Noveck, I.A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14(4), 347–375.
- Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30(2), 191–205.

4

**An investigation of
the violation tolerance hypothesis**

Katrijn Pipijn
Walter Schaeken

Abstract

Children and adults seem to interpret scalar implicatures in a different way. Research with a scalar response measure however has shown that children and adults are more alike than a binary response measure reveals. The pragmatic tolerance hypothesis has been proposed to explain this phenomenon but unfortunately this theory does not explain how children interpret reversed scalars. We therefore propose a broader violation tolerance hypothesis and claim that children have an overall higher tolerance for violations, independent of whether they are logical or pragmatic violations, than adults. We used a testing paradigm in which adults had to rate scalar implicatures, reversed scalars and control items, which enabled us to distinguish between the two theories. Results of the study did not favor the violation tolerance hypothesis over the pragmatic tolerance hypothesis.

4.1. Introduction

Verbal communication is one of the cornerstones of human civilization and is one of the most important aspects of our daily lives. Various aspects of linguistics have been studied for centuries yet one particular aspect has gained popularity in recent years: scalar implicatures. Consider the following example:

Some of the players in the World Cup final played well.

When a person utters this sentence, you can deduce many different things about it. For example, you might assume that the person watched the final game of the World Cup. You will probably also deduce that the person thinks that not all the players played well and that there is at least one player that did play well and one that didn't. However, it is also possible that the person did not actually watch the game and only read some articles about it, in which case he might have read something about certain players' performances, but not about others. In this case it is possible that he used 'some of the players' because he does not have information about all the players. So it is clear that there are two alternative interpretations of this sentences that are both equally correct.

Some and possibly all players in the World Cup final played well.

Some but not all players in the World Cup final played well.

It is up to the listener to decide which of these two interpretations he or she will adhere to. To make this decision, the listener will use all his prior knowledge and context information. The listener knows for example

that the speaker despises soccer and definitely did not see the final. Or perhaps the sentence is uttered in a bar right after both the speaker and listener watched the game together attentively. The listener will incorporate all this information while interpreting the sentences. Incorporating all this extra context information is what we call pragmatics. It is obvious that interpreting this sentence is much more than processing all the semantic information the correct way, pragmatics is clearly another important part. In this particular example, we saw an illustration of a scalar implicature. Grice (1991) introduced a theory that explains these pragmatic circumstances. Grice believes that people obey certain 'rules' when communicating, which he calls 'Maxims'. For example there is the Maxim of Quantity: Make your contribution to a conversation as informative as required. For our example this means that the speaker needs to give all the information he has and be as informative as he can be. However, he also needs to obey the Maxim of Quality which limits him to only give information that he knows is true. For our example this means that our speaker cannot use the term *all* if he is not absolutely sure that it is correct. If we assume that the speaker is applying both these maxims, then the initial statement needs to be interpreted with the '*not all*' meaning. The correct interpretation is *implicated* by the speaker.

The reason why this type of implicature is called a 'scalar' implicature goes back to the work by Horn (1984, 1989). According to Horn, the derivation of this type of implicatures is based on so-called Horn scales. Examples of these of scales are <*some, all*>, <*possible, certain*> and <*warm, hot*>. Typical about these types of scales is that each word further on the scale is entailed by the weaker previous word on the scale. The word *all* is stronger than the word *some* and it is entailed by it. This works only in one direction. *All* does not entail *some*. The word

some literally means 'at least one and possibly all' so *all* is entailed in the definition. When a person uses a less informative word of a scale in a sentence, it is probably the case that the stronger word is not the case. If it were, and the speaker obeys Grice's Maxims then he would have used the stronger word. Therefore, the best way to interpret the weaker term would be in the meaning that excludes the stronger term, which is also called the pragmatic interpretation. These interpretations go far beyond the semantic meaning of what is said. Pragmatics plays a large role and communication is clearly not only about what is said, but maybe even more about what is not said.

Research on the development and the general mechanics behind scalar implicatures has really lifted off after a study conducted by Noveck (2001). Noveck conducted a developmental study in which children and adults had to indicate whether they agreed or disagreed with statements. The critical statement in the study was an underinformative item that could be interpreted either in a logical fashion or in a pragmatic fashion. In his third experiment he copied a study conducted by Smith (1980) in which sentences like 'Some giraffes have long necks' are presented to participants. There are two ways to interpret this sentence, the logical way (Some and possible all giraffes have long necks) and the pragmatic way (Some but not all giraffes have long necks). The results showed a clear discrepancy between children and adults in how they interpret this sentence, with up to 87 percent of children accepting this statement while only 47 percents of adults do so. Adults appear to be more likely than children to enrich the interpretation of the world *some* to *some and not all* and reject these types of statements.

More recently, Noveck and Sperber (2007) wrote a review on theoretical and developmental aspects of scalar implicatures. The devel-

opmental studies discussed in the review (Guasti et al., 2005; Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004) point into the direction that children are less capable of making pragmatic inferences than adults. When the circumstances are right, however, children's performance can be improved, for example when they are given sufficient training. Still, a significant difference with adults remains. It seems that making pragmatic inferences requires more effortful processing from children than it does from adults and when confronted with a weaker scalar term, children will be more likely to give a logical interpretation of the term as opposed to a pragmatic one.

In 2011, Katsos and Bishop were able to make the differences between children and adults disappear. In an earlier study, Katsos and Smith (2010) found that when children had to rate scalar inferences on a 5-point Likert-scale, a majority of the children chose the middle options of the scale, expressing their sensitivity to the underinformative nature of the scalar impressions. In the 2011 study (Katsos & Bishop, 2011), a comparison with adults was made. All participants in the study were shown a story that was narrated by a fictional character. The participants then had to rate a statement uttered by the fictional character. This statement would be optimal, false or underinformative. In a first experiment, the participants had to give a binary response. In the second experiment, they had to give a scalar response, on a 3-point Likert-scale. In the first experiment with the binary responses, the researchers found a significant difference between children and adults on the underinformative responses. In the second experiment however, this significant difference between the children and the adults disappeared. These studies clearly show that children are sensitive to the ambiguous nature of scalar implicatures but that a two-alternative forced choice paradigm, which is used

in most common scalar implicatures studies, is not adequate to expose this sensitivity.

We were able to replicate these findings. In the first experiment of one of our studies (Pipijn & Schaeken, 2012) we looked into the effects of scalar response options on the interpretation of underinformative items with three different tasks. We replicated the results found by Katsos and colleagues; we did not find a significant difference between children and adults when interpreting scalar implicatures with a Likert-scale. We used three different tasks to play with the task difficulty. The effect of the scalar response options seemed to be robust and did not alter with the task difficulties. In the second experiment we conducted a within-subjects experiment in which each participant had to rate sentences both binary and on a scale. We tested both children and adults, on two different tasks that varied in difficulty level. Consistent with the literature we found that children and adults interpret scalar implicatures differently when they have to give binary responses. This difference disappears when they have to give the answers on a scale. Adults seem not to be influenced by the task difficulty but children are. Especially when they have to give scalar ratings of underinformative items, a difficult task leads them to be more logical than with an easier task. This effect is consistent with what we could expect from previous research and what we previously said; making scalar implicatures requires more cognitive effort from children than it does from adults and this becomes clear when you compare performances on a difficult and an easier task.

These studies and the studies by Katsos and colleagues (Katsos & Bishop, 2011; Katsos, Roqueta, Estevan, & Cummins, 2011; Katsos & Smith, 2010) clearly show that children do have the required sensitivity to interpret weak scalar terms in a pragmatic fashion. Yet, it remains the

question why this sensitivity does not show in classic binary response paradigms. Katsos and colleagues came up with a theory that explains these findings very well, called the pragmatic tolerance hypothesis (Katsos & Bishop, 2011).

The starting point of this theory is that in order to pragmatically reject underinformative sentences, a participant does not only need to have the sensitivity to notice the under-informativeness of a scalar term, the participant also needs to classify this use of scalar terms as a violation. We call these two steps pragmatic competence and pragmatic tolerance. It is possible that a person notices this underinformative use, but does not classify it as a violation. Even more, it is possible that the person classifies it as a violation, but not as a violation grave enough to outright reject it. So when a person accepts the underinformative use of a weak scalar term in a binary task paradigm, there is no way of knowing whether this acceptance is due to insensitivity or due to a classifying decision. The use of Likert-scales gives a solution to this problem and the previously mentioned studies show it is not an otiose remedy. When participants are confronted with this scale, it gives them the perfect tool to express this sensitivity.

There are two possible types of violations in the items that we used: logical ones and pragmatic ones. The logical ones are straightforward and lie in the semantics of the scalar terms. When a strong term with a delimited meaning is used to describe a situation that is not entailed in the meaning of the word, it is incorrect in any way you look at it. For example saying 'all cats have wings' is logically incorrect because *no* cats have wings. The pragmatic violations are more vague and lie in the wide meanings of the weaker scalar terms. When a weak term is used to describe a stronger situation that might, but does not have to, be entailed

in the weaker term, there is not a clear-cut right or wrong answer. For example saying 'some cats have whiskers' when in fact *all* cats have whiskers. Instead there are two ways to look at it and it depends on context or personal preference to interpret the scalar term and the situation. For example when the whiskers of a cat are cut, a person can see this as a cat without whiskers while another person might not see it that way because the cat was born with them in the first place. These two variables give rise to the pragmatic tolerance hypothesis. It is up to each participant to draw the line as to what is or is not acceptable to him. Katsos and colleagues call this personal threshold for acceptable and unacceptable violations a tolerance. They believe this tolerance does not apply to logical violations because there are no vague interpretations possible. The tolerance is only applicable on pragmatic violations. The location of the threshold will be different for adults and children. For children, who still have a limited understanding of language and especially pragmatics, this threshold is fairly high and they are tolerable. Adults on the other hand are more experienced thinkers and linguists and therefore their threshold will be stricter and less tolerable. When both children and adults are confronted with a pragmatic violation like an underinformative use of a weak scalar term, adults will therefore classify this violation as too grave a violation and reject it while children might not. This leads to different response patterns for children and adults when they have to respond binary. Even though they might both experience the conflict between the logical and the pragmatic interpretation in the same way, their individual thresholds will lead to opposite response patterns. When they can answer on a scale though, both children and adults have a lot more options to express their sensitivity. They will both express their conflicting feelings by using the middle options on the scale. By doing this, they can

acknowledge that there are two conflicting interpretations but that one is not necessarily better than the other.

This hypothesis by Katsos and colleagues is elegant and easily explains for all the scalar data they found. It does not however explain all the data that were found in our previous experiments (Pipijn & Schaeken, 2012; Pipijn & Schaeken, 2013). In these two studies, we did not only investigate scalar implicatures, we also investigated something we call reversed scalars. Reversed scalars are what most other studies consider being one of the various control items. It is the reverse of a normal scalar implicature. More specifically, a reversed scalar is an item in which a strong scalar term is used to describe a situation in which a weaker term would have been appropriate. An example of this would be when someone says 'All dogs are brown'. The stronger term *all* is used but the weaker term *some* would have been correct because there are also black dogs for example. In our studies we found that for children, these items are equally conflicting as scalar implicatures. We found that children rated reversed scalars different from other control items in that they used the middle of the scale to rate them. When children had to respond binary, they were clearly aware that these items were incorrect, yet the scalar responses do point in the direction of some sort of ambiguity. Adults also clearly rejected these items when responding binary. When responding with scalar responses, they also showed a small reluctance to outright reject these items. Their average scores were lower than those of children, but still, they did show some minor uncertainty about these items.

These findings cannot be explained by the pragmatic tolerance hypothesis. The errors made by the participants were logical semantic errors and this type of violation does not fall under the pragmatic tolerance hypothesis. Therefore we propose a revision of the hypothesis and

we advocate a more general violation tolerance hypothesis. We believe that not only pragmatic violations are judged by a personal violation tolerance threshold, but also all sorts of violations are proportioned to a tolerance threshold, including logical violations. For adults the gravity of a logical violation will be much larger than the gravity of a pragmatic violation. For children, the difference between the two will be much smaller. While a pragmatic violation that is rated at the middle of the scale might be either accepted or rejected binary, a logical violation that is rated at the middle of the scale, will be rejected without a doubt. For adults the logical violation is so grave to begin with, that even on a scale, it raises little doubt.

In this paper we will attempt to explore this hypothesis. To do this, we used a paradigm that enables us to vary the magnitude of a violation, which should allow us to explore the violation tolerances. We got inspiration from a study conducted by van Tiel (2014). In a study on embedded scalars and typicality, van Tiel (2014) used a paradigm that is very suited to test our hypothesis. We made some minor modifications to the paradigm to make a comparison with our previous studies possible. In the experiment we will use a set of pictures with black and white dots, in which we will vary the ratios of black and white dots. Then we will ask participants whether they agree with statements about the dots. These statements will contain the words *all*, *some* and *none*. Participants will give their responses on 5-point Likert-scales. The largest difference between our study and the study by Van Tiel is that we also included the word *none*. These items are crucial to make a distinction between the pragmatic tolerance theory and our violation tolerance theory, which we will explain further on.

Purely based on semantic theories, we would expect participants to use the top rating of the scale if the word *all* is combined with a situation in which all the dots are black and if the word *none* is combined with a situation with no black dots. For all other combinations of these words and dots ratios, we would expect them to pick the lowest option on the scale. For the sentence with the word *some* we expect them to pick the lowest point on the scale when there are no black dots and we expect them to pick the top of the scale for all other black and white dots ratios, except for the situation in which all the dots are black. This particular item is the scalar implicature because although the word *all* would be better suited, the *some* term is not necessarily incorrect. Therefore all options on the Likert-scale are acceptable for this particular item.

If we would base our predictions on our violation tolerance hypothesis however, they would slightly alter. For the items in which the *all* term is used, we expect a small positive slope in mean ratings as the ratio of black and white dots goes up. For the sentences with the word *none* we expect the same but in the other direction, namely, a small negative slope in mean ratings when the ratio of black and white dots goes up. We expect the slope for the *none* sentences to be less steep than the slope of the *all* sentences. The reason for this is the one-directionality of the Horn scales. The word *some* might entail the word *all* but this does not work in the direction of the word *none*. This makes the semantic distance between the words *some* and *all* smaller than the semantic distance between *some* and *none*. Therefore, we think that the incorrect use of the latter pair will be a graver violation than the incorrect use of the former pair. All the items in which we use the word *all* but not all the dots are black, are examples of reversed scalars. The large difference between our violation tolerance hypothesis and the pragmatic tolerance

hypothesis lies in these reversed scalars. The violation tolerance hypothesis sees these reversed scalars as logical errors that might be acceptable to some people or in some situations. Therefore we expect the positive slope. Within the pragmatic tolerance hypothesis, these errors are never acceptable. If this were true, there will be no slope in the data of these items. This difference between these two theories is also why we included the none items. The violation tolerance predicts a slope in the data for both these sentences, but the slope of the *all* items should be larger. The pragmatic tolerance hypothesis or semantic rules do not predict a slope for either of these sentences, so no significant difference between the two slopes should be found.

For the *some* items, we expect that the mean ratings will not be pushing limits for all the ratios, instead we expect there to be a gradual increase and decrease over the ratios. We expect the peak of the ratings to be at what each participant perceives at the ideal ratio for the word *some*. Based on typicality research of the word *some* (Begg, 1987), we know that 'the preferred meaning for some is 'less than half'' (p.62). Several other authors have made similar conclusions (Borges & Sawyers, 1974; Newstead, Pollard, & Riezebos, 1987). For the situation in which all the dots are black, we expect participants to use the middle of the scale, consistent with the previously discussed scalar implicature literature. Both the violation tolerance hypothesis and the pragmatic tolerance hypothesis predict the same data pattern for the *some* items.

In our experiment, we varied the total amount of dots used. Newstead, Pollard, and Riezebos (1987) showed how varying the set size influences how people will interpret scalar term. It seemed that a larger set size leads to interpretations that signify a smaller proportion of that set size than when a smaller set size is used.

4.2. Method

Fifty-six adults were recruited for the experiment, of which 22 received course credit in exchange for their participation, the others volunteered. Two participants were excluded because they did not finish the experiment. The experiment was programmed in Qualtrics, an online survey platform.

For each item, the participant was presented with a picture and an accompanying sentence about the picture. An example of an item is presented in Figure 1.

All dots are black

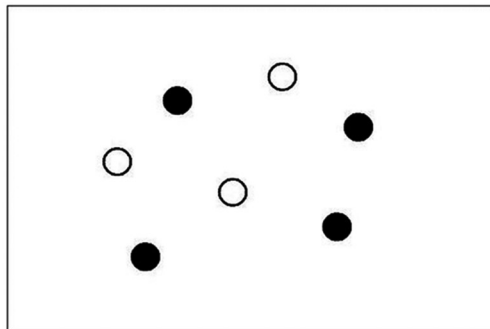


Figure 1. Example of an item with sentence and picture.

In the picture there were always a number of dots that were either black or white. We had one between-subject variable: total number of dots, with two levels. Half of the participants got pictures with 7 dots, the other half got pictures with 35 dots. The participants were assigned to a condition randomly. The ratio between black and white dots was varied

and each participant received each possible ratio (within-subject variable ratio). For the second number of dots condition with the 35 dots, the number of black dots was always a plural of five (0, 5, 10, 15, 20, 25, 30, 35), leading to the exact same ratios of black/white dots as the first condition. The picture was also accompanied with one of three describing sentences ('All the dots are black', 'Some the dots are black', 'None of the dots are black'). Each picture was presented to the participants three times, so every combination of ratio – scalar term was rated. The items were presented to the participants in a random order. Participants were asked how well the sentence described the picture. They had to give their answer on a 5-point scale, ranging from not good to very good.

Items in which all the dots were black and the sentence 'Some dots are black' was given, is considered a scalar implicature. The word *some* is underinformative in this example and the word *all* would have been more optimal.

4.3. Results

In the data analysis, we will compare the observed results to what we would expect them to be when participants follow semantic rules. We will call these data the estimated data. We expect that for the trials in which the *all* sentence is used, participants will pick the lowest option on the scale, except for the situations in which all the dots are black, in which case we expect them to pick the highest option on the scale. For the trials in which the pictures are combined with the sentence 'No dots are black', semantics predict the opposite pattern. For the *some* situations, we expect the lowest score for the picture with no black dots and a maximum score for the middle options. For the combination of the *some*

sentence and the picture in which all the dots are black, any score between 1 and 5 is valid. This particular item is the scalar implicature. A logical response on this item would be a 5 and a pragmatic response would be 1. However, because the participants are given a scale to answer, we expect them to use the middle of the scale to express this conflict.

We ran a repeated measures design with scalar term and ratio of dots as within-subjects variables and total number of dots as a between-subjects variable. We corrected our p -values with the Greenhouse-Geisser method (p_{GG}) because the assumption of sphericity was violated. We found main effects for scalar term ($F(2, 98) = 173.85, p_{GG} < .001$) and ratio of dots ($F(7, 343) = 27.34, p_{GG} < .001$). There was no main effect for total number of dots. The scalar term interacted with ratio of dots ($F(14, 686) = 230.17, p_{GG} < .001$), but not with total number of dots. Ratio of dots interacted with total number of dots ($F(7, 343) = 5.01, p_{GG} < .001$), and there was also a three-way interaction with scalar term ($F(14, 686) = 4.96, p_{GG} = .001$).

We ran paired samples t -tests between the observed data and the estimated data, for all three scalar term types. We found that for the *all* and the *some* items there was a significant difference between the observed data and the data we would expect based on semantics (*all*: $t(431) = 5.58, p < .001$; *some*: $t(428) = 14.87, p < .001$). For the *none* items there was no significant difference between the observed and the estimated data ($t(431) = 1.62, p = .11$). In Figure 2, 3 and 4, we clearly see how, especially for the *all* and the *some* items, there is a difference between the observed and the estimated data. The *none* data lies much closer to the estimated data.

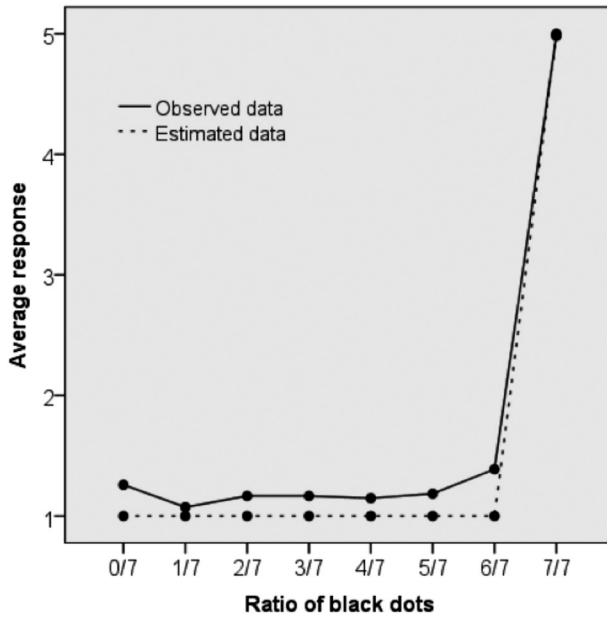


Figure 2. Average responses for 'all' items for the different dot ratios.

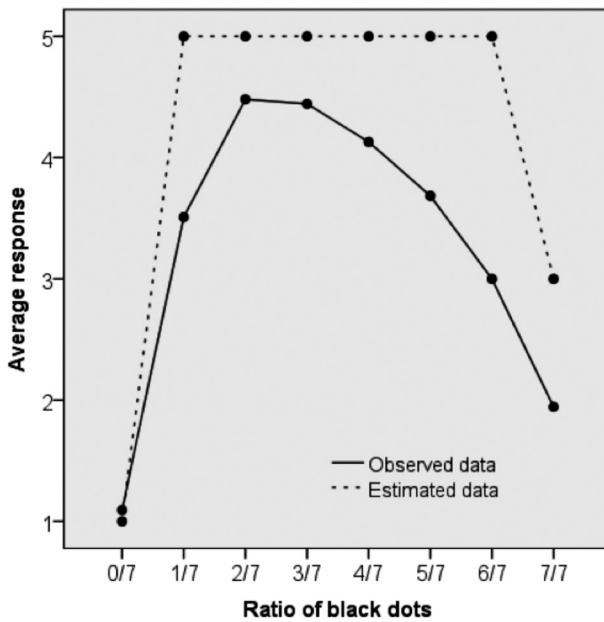


Figure 3. Average responses for 'some' items for the different dot ratios.

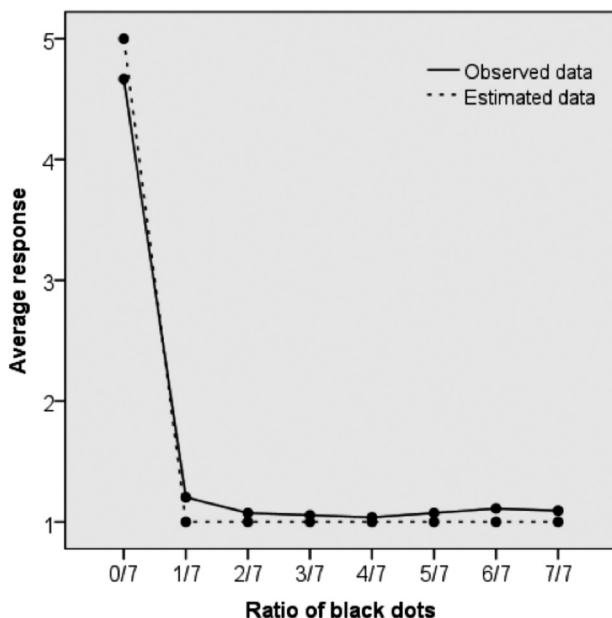


Figure 4. Average responses for 'none' items for the different dot ratios.

Next, we want to check whether the answering patterns for the *all* and *none* items increase and decrease linearly, as we would expect according to the violation tolerance hypothesis. We find that the ratio of dots predicts the *all* items linearly ($\beta = .33, t(430) = 13.24, p < .001$). We found the same results for the *none* items ($\beta = -.30, t(430) = 13.46, p < .001$). These significant results can probably be explained by the last and first ratio of the *all* and *none* items, respectively. Therefore, we will run these analyses again but without the 7/7 ratio of the *all* items and the 0/7 ratio of the *none* items. Now, we do not find a significant linear trend for the *all* items ($\beta = .021, t(376) = 1.22, p = .22$), nor did we find one for the *none* items ($\beta = -.009, t(376) = .84, p = .40$). Lastly, we checked whether the slope of either of these variables was larger than the other,

but again we did not find an effect ($\beta = -.016$, $t(376) = .74$, $p = .46$). These results indicate that the differences between the *all* and *none* sentences we found with the *t*-tests are not explained by a linear trend or the lack thereof in either one of the variables. It seems that these results do not support the violation tolerance hypothesis.

For the *some* sentences we clearly see in Figure 3 that the observed and the estimated data are different from each other, except for the 0/7 ratio. The results for this item type do point in the direction of the violation tolerance hypotheses. The average response for the 7/7 ratio is, like we expected for a scalar implicature, somewhere in between the lowest and the highest score ($M = 1,94$; $SD = 1,27$). For the middle ratios, there seems to take on the shape of a right-skewed parabola. It seems that the optimal ratio for the world *some*, lies around 2/7. However, when we separate these data for the two total number of dots conditions (Figure 5); we see that for the high total number of dots, the graph is even more skewed. It seems that it is not so much the ratio of black and white dots that is important, but more the actual number of black dots. This is probably the explanation for the interaction effect between the ratio of dots and the total number of dots.

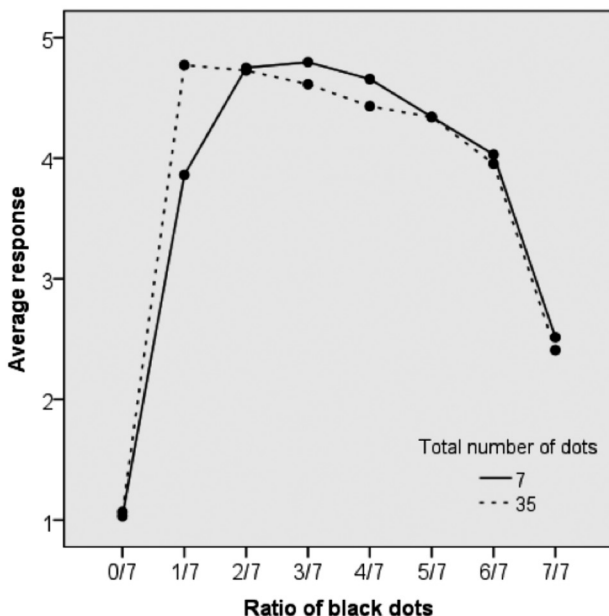


Figure 5. Average responses for the 'some' items for the two total number of dots condition.

4.4. Discussion

We explained that there are two possible patterns that our results could take on. The pattern that we would expect if participants rated the stimuli purely based on semantic rules, and the pattern we expect based on our violation tolerance hypothesis. Our results do not show a better fit for one pattern or the other. We found that the stimuli that contained the words *all* and *none* follow the pattern based on semantic rules. Participants chose the highest rating for the optimal situations, which are when all the dots are black in the *all* sentences and when there are no black dots for the *none* sentences. For all the other situations, participants chose the lowest ratings. Although we found a difference between the

predicted pattern and the actual data for the *all* sentences, this pattern could not be explained by any further analyses. We also did not find a significant difference between the *all* and the *none* sentences. This is consistent with the expected pattern based on semantic rules and the pragmatic tolerance hypothesis, but not with the violation tolerance hypothesis. In the violation tolerance hypothesis we expected there to be a difference based on the semantic characteristics between *some* and *all* which are not there between *some* and *none*.

The *some* items did not follow the semantic pattern. Like we expected, the ratings for the different ratios varied gradually. As expected, the rating was the lowest when there were no black dots, it increased gradually over the items, with the highest rating for a ratio that was less than half, and then it decreased again. For the item in which all the dots were black, the scalar implicature, we found an average score in between the highest and the lowest. These results are in line with the violation tolerance hypothesis. There are however also congruent with the pragmatic tolerance hypothesis. The *some* items are not the items that would have enabled us to distinguish between the two theories, only the *all* items and a comparison between the *all* and *none* items could have made this distinction. Like we said in the previous paragraph, our data did not show a difference between the *all* and *none* items and we were not able to confirm the violation tolerance hypothesis.

We did not find a main effect of total number of dots, nor did we find an interaction with scalar term. We did however find an interaction with ratio of dots and a three-way interaction. The three-way interaction is explained by the fact that the total number of dots is only significant in the *some* condition. We saw in Figure 3 that the pattern was right-skewed, with a peak around 2/7. When we looked at the two total

number of dots conditions separately (Figure 5), we saw that there is in fact a difference between these two conditions. The pattern for the high number of total dots is even more skewed than the condition with seven dots. The peak here lies at about $1/7$ ($5/35$). This number is much lower than what we would expect based on the literature, which states that it should be 'less than half'. This makes us wonder whether it is not as much the proportional number of black dots that matters but more the actual number of dots that is important. Also, it is possible that the language aspect is an important factor in explaining our divergent results. Our previous studies with both adults and children (Pipijn & Schaeken, 2012; Pipijn & Schaeken, 2013), complemented with other studies that were conducted in Dutch (De Neys & Schaeken, 2007; Geurts & Pouscoulous, 2009) show much higher rates of pragmatic reasoning than studies in other languages (Bott & Noveck, 2004; Noveck & Posada, 2003). The interpretation of the word *some* might very well differ between languages. In that case, the optimal number or proportion for the word *some* could be lower in Dutch than it is in other languages. A study with either a much higher number of dots than 35 or with 35 dots in which each possible ratio of dots is presented, or a language comparison study, would bring more clarity on this issue.

Like we already mentioned, our results pointed in the direction of the pragmatic tolerance hypothesis and not to the violation tolerance hypothesis. In light of this, there are some general considerations we would like to point out. First of all, we are very surprised that we were not able to replicate Van Tiel (2014) with our experiment. We did not find the steep positive slope in the *all* data that was present in the Van Tiel study. Our paradigm was very similar to his so we expected the same patterns, but we did not find them. The pattern for the *some* sen-

tences was fairly similar to ours with the only difference being that in the Van Tiel study the highest rating was found just below half of the dots, a much higher ratio than in our study and congruent with the literature. The *all* sentences pattern was different. The pattern found in Van Tiel was just about what we expected on the basis of the violation tolerance hypothesis. Van Tiel found a gradual increase in ratings. Van Tiel did not include the *none* condition in his study. It is unclear to us why we were not able to replicate these results. The adjustments we made to the paradigm were only minor. First of all, we changed the total number of dots. Van Tiel used ten dots in every condition. Secondly, he used a 7-point rating scale, as opposed to the 5-point scale we used. These adjustments seem only minor but perhaps they had a larger impact than we expected them to have.

There are two other factors that might explain these differences. First of all, Van Tiel did his study in English, while we did ours in Dutch. As previously mentioned, there are language differences in the interpretation of scalar implicatures. Dutch participants have already proven to be more pragmatic than English or French participants. The direction of this language difference only makes the situation more confusing though. If Dutch participants are in fact less logical and thus less concerned with the literal meaning of words, then you would expect them to be more open to semantic errors like reversed scalars. On the other hand, if you look at it from Grice's Maxims it is possible that Dutch participants give a higher priority to the Maxim of Quality which states that all given information has to be true. This Maxim does not leave room for ambiguous interpretation of words and this would apply to both scalar implicatures, which would lead them to be pragmatic, as to reversed scalars (which would lead them to explicitly reject them). A second factor that was al-

tered in our study and that could make a difference, is that Van Tiel used the word *every*. In our study, we used the Dutch word *alle*, which closest translation in English would be *all*. The closest Dutch translations of the word *every* however, would be *iedere* or *elke*. While all these words lie very close to each other, both in English as in Dutch, it is still possible that they do lead to small interpretational differences that influenced our results.

A second large discussion topic about our study is that this study was conducted with adults. Results would probably have been very different if we had tested children and we strongly believe that a study with children should be conducted in the future. Our previous studies have shown some clear differences between children and adults, especially for the reversed scalar items. The reversed scalar items in this study would have been all the items in which not all the dots were black, combined with an *all* sentence. In previous scalar implicature research on adults, the interesting reversed scalar phenomenon was not found. For example, in a study by Bott and Noveck (2004), the reversed scalar items were never found to be any different from other control items. Not even in the reaction time data was there any sign of participants treating these items differently. If adult participants did experience some sort of ambiguity about these items, this could be manifest in the reaction time data, regardless of their final response. The data in the study by Noveck (2001) already showed a small indication in this direction. He found that children of 7 to 8 years old were less accurate on these items than they were on other control items, but children of the ages of 10 and 11 did not show this effect any more. We do not know whether this difference with the other control items was significant. Even if adults do have some ambiguous feelings about this type of item, they might be too small and

our experiment not sensitive enough to pick them up. In this light we want to recall the significant t -test between the estimated data and the observed data for the *all* sentences. Perhaps this result is a hint towards the effect we are speaking of. Yet it remains unclear why this experiment did not pick up on this effect while our previous experiments did (Pipijn & Schaeken, 2012). Perhaps the nature of our stimuli is to blame. In our previous experiments we used mostly drawings as stimuli while we used the more monotonous dots in this experiment. Scalar implicature research is known to be sensitive to changes in task features; perhaps this change was too big.

A logical next step would be to test children, perhaps even preschoolers, with the same paradigm used in this study. Our previous experiments have showed that the reversed scalars effect is a lot more robust in children, especially in preschoolers (Pipijn & Schaeken, 2013). We expect that preschoolers and children are a lot less semantically correct than adults about all items. We expect them to confirm the violation tolerance hypothesis in a way adults clearly could not. We would expect a similar pattern for the *some* items as adults, but we also expect that for the *all* and the *none* items there is a small but significant slope in the response pattern. Even though our experiment was not able to confirm the violation tolerance hypothesis yet, the results from our previous studies clearly indicate that there is more to the processing of the words *some* and *all*. These results would be able to unambiguously confirm or reject the violation tolerance hypothesis.

References

- Begg, I. (1987). Some. *Canadian Journal of Psychology*, 41(1), 62.
- Borges, M. A., & Sawyers, B. K. (1974). Common verbal quantifiers: Usage and interpretation. *Journal of Experimental Psychology*, 102(2), 335.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3), 437–457.
- De Neys, W., & Schaeken, W. (2007). When People Are More Logical Under Cognitive Load. *Experimental Psychology*, 54(2), 128–133.
- Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics and Pragmatics*, 2, 4–1.
- Grice, H. P. (1991). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (Ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Horn, L. (1989). *A natural history of negation*. Chicago: Chicago University Press.
- Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.
- Katsos, N., Roqueta, C. A., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119(1), 43–57.

- Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness. In K. Franich, K.M. Iserman, L.L. Keil (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*. Somerville, MA: Cascadilla Press.
- Pipijn, K. & Schaeken, W. (2012). Children and pragmatic implicatures: A test of the pragmatic tolerance hypothesis with different tasks.
- Pipijn, K. & Schaeken, W. (2013). Is it Tolerance or Pragmatic Tolerance? The Pragmatic Tolerance Hypothesis in preschoolers.
- Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics*, 18(3), 178–182.
- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2), 165–188.
- Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2), 203–210.
- Noveck, I. A., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of 'Scalar Inferences'. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 184-212). Palgrave: Basingstoke.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253–282.
- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12(1), 71–82.
- Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30(2), 191–205.

- Teresa Guasti, M., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, 20(5), 667–696.
- Van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, 31(2), 147–177.

5

How scales are influenced by scales

Katrijn Pipijn
Leen Janssens
Lise Vaes
Walter Schaeken

Abstract

Many studies on scalar implicatures have focused on the flagship example of <some, all>. This study joins a growing amount of research that addresses the issue that there is in fact large scalar diversity and not all the different scales can just be generalized. In this study, we conducted two experiments with several different scalar implicatures. In the first experiment we tested adults, in the second experiment we tested preschoolers. Not only were participants asked to judge scalar implicatures in a binary fashion, they were also asked to rate them on a Likert-scale. In the experiment with the adults, we varied the amount of available cognitive resources by adding a secondary task. Our study shows that there clearly is scalar diversity between quantifiers and gradable adjectives. This diversity is further enlarged by the different response methods. No effect of available cognitive resources was found. Our results show that the scope of scalar implicature research needs to be broadened to give a realistic representation of pragmatic communication.

5.1. Introduction

The complexity of human conversations provides endless inspiration for research and discussion. One of the discussion topics is the nature of scalar implicatures. In a normal conversation, people utter words for which the meaning extends far beyond the literal meaning and often, the listener comprehends that intended meaning. Take into consideration the following example:

- (1) Some people find the Belgian governmental structure complicated.
- (2) Not all people find the Belgian governmental structure complicated.
- (3) All people find the Belgian governmental structure complicated.

It seems obvious that sentence (1) will be interpreted as sentence (2) most of the time, instead of sentence (3). However, according to the logical semantic meaning of the word *some*, which is *some and possibly all*, sentence (3) is also an accurate interpretation of sentence (1). Then why do most people prefer the interpretation in sentence (2)? The reason for this lies in the pragmatic interpretation of *some* as *some and not all*. Horn (1984) sees this example as a variety of the conversational implicature. When a person is confronted with a sentence like (1), his or her initial interpretation will be *some and possibly all*. Later on, the listener may change this interpretation to the pragmatic one and make the scalar implicature. If the speaker had the intent to refer to *all* the people, then surely he would have done so by saying *all* and not *some*. Therefore, the

all interpretation of *some* is not optimal and the *not all* interpretation is far more likely. This argumentation is based on what Grice (1989) calls the cooperative principle. People are to be as informative as possible in conversations. When they use a weaker term instead of a stronger one, it must be because the stronger one is not suitable. There are however different opinions on how this pragmatic meaning comes to be. As opposed to Horn (1972), Levinson (2000) believes that not the logical meaning, but the pragmatic meaning is default in human thinking. This pragmatic interpretation might then be cancelled when the situation calls for it.

The words *some* and *all* can be put on a scale going from weak to strong. This is why we call this pragmatic phenomenon a scalar implicature. Of course, *<some, all>* is not the only example of a scale that is used in scalar implicature research. Another frequently used scale is the *<or, and>* scale, where *and* is entailed by the weaker term *or*. Most studies in this research area have used these two scales, *<some, all>* and *<or, and>*, as the prototypical examples of scalar implicatures and the results of these studies have been generalized to all scalar implicatures. Other possible examples of scales are verbs (*<might, must>*, *<like, love>*), adjectives (*<good, excellent>*, *<warm, hot>*), adverbs (*<sometimes, always>*) and nouns (*<mammal, dog>*).

Until recently, this alleged uniformity of the processes underlying all scalar implicatures had not been questioned. In 2009, Doran, Baker, McNabb, Larson, and Ward started to explore this widely accepted assumption. In their study they compared quantificational items like the classic *<some, all>* but also others like *<possibly, definitely>*, to other scales as for instance cardinal numbers (e.g., *<2, 3, 4>*), ranked orderings (e.g., *<beginner, intermediate, advanced>*) and gradable adjectives (e.g.,

<*warm, hot*>). They found that there is significant variability between the rates of pragmatic answers these different scalar terms elicit: Some scalar inferences are consistently interpreted in a pragmatic way while others are mostly interpreted logical.

Other evidence against this uniformity principle between scalar implicatures can be found in a survey of ten experiments, which was carried out by Geurts (2010, 98-99). A clear conclusion of this survey was that for disjunction sentences (containing *or*), the mean rate of scalar implicatures was much lower than for the sentences containing *some*: 35% against 56.5%.

All the studies discussed above used a sentence verification task. A straightforward example of this paradigm can be found in the study by Bott and Noveck (2004, Experiment 3). They presented stimuli of the following example:

(4) Some dogs are mammals.

Participants were subsequently asked whether they believed this sentence was true or not. Typical about this type of experiment is that participants have to judge the items with their own intrinsic knowledge. When confronted with items of this type, participants derived scalar inferences for 59% of the items.

Doran et al. (2009) used a different approach. Participants were instructed to judge scalar implicatures, but from the perspective of Literal Lucy, a literal-minded character. They were asked whether Literal Lucy would agree with a certain utterance. Therefore, participants were not actually giving their own opinion. By asking for the opinion of Literal Lucy, participants are already guided towards an opinion that might be more

logical than their own. It is clear how the difference in approach between Bott and Noveck (2004) and Doran et al. (2009) can lead to differences in pragmatic response patterns.

The work done by Van Tiel, Miltenburg, Zevakhina, and Geurts (submitted) builds further on the work by Doran et al. (2009). Taking into account some of the issues they had with the work done by Doran et al., they constructed experiments to further explore the non-uniformity between different scales. First of *all*, they changed the task paradigm to an inference paradigm. In Geurts and Pouscoulous (2009), it was shown that an inference paradigm is much more suited than a verification task to explore the rich variety of scalar inferences. Van Tiel et al. conducted several experiments, *all* with the purpose of finding parameters that can change the scalar inference ratings. They included several different types of scales or word classes: quantifiers, modal expressions, gradable adjectives and verbs and they played with various characteristics of the scales. They looked at open versus closed scales (for example, *all* is the end point of the <*some, all*> scale, which is thus a closed scale, while for example *cold* is not the end point in <*cool, cold*>). They also gave adjectives a richer context, for example <she is attractive but not stunning>. Other parameters they looked at were focus (whether the scalar term was the focus of an utterance or not), the word frequency of the two scalar terms in our everyday language, the strength of association between a weaker and a stronger scalar term, and the semantic distance between a weaker and a stronger scalar term. In their first experiment, they found large difference in rates of scalar inferences, namely between 4% and 100%. They found very high scalar inference ratings for existential quantifiers, which are closed scales, of more than 90%. Results show a clear tendency to interpret these implicatures pragmatically. This is a clear difference with

the results found by Bott and Noveck (2004). They found a significant difference between closed and open scales, with closed scales leading to more scalar inferences. Similarly to Doran et al. (2009), they additionally found that giving the adjectives a richer context leads to more scalar inferences. Lastly, they found that word class and semantic distance have a significant effect on the rate of pragmatic responses. However, they did not find any effects of focus, word frequency, or strength of association between stronger and weaker terms. Their study clearly demonstrated that different types of scales are not all the same and we cannot use one type as the prototypical type. Especially the *some, all* scale cannot be used as the prototypical example, because it triggers unusually high levels of pragmatic answers.

5.2. Experiment 1

A first goal of the present study is to assess the effect of another parameter on the different scales, namely response measure. In 2010, Katsos and Smith did research on scalar inferences with children and adults, while changing the response method. They introduced Likert-scales in the scalar inferences debate. Instead of just saying whether they agreed with a certain utterance or not, participants now had to indicate to what extent they agreed with the utterance. Both adults and children accepted correct control statements and rejected the false ones by using the extremes of the scale. However, for the scalar inferences, both children and adults frequently picked the middle of the scale. There was no significant difference between children and adults when it came to these scalar inferences. This is in strong contrast with other studies on scalar inferences with children. For example, Noveck (2001) found large

significant differences between children and adults. Katsos and colleagues replicated the effect of response type in other studies, thus it seems to be a robust finding (Katsos & Bishop, 2011; Katsos, Roqueta, Estevan, & Cummins, 2011). In the present study, we are therefore in the first place interested how this Likert-scale would influence other types of scales.

A second parameter that will be explored in the present study is the effect of load on different scales. In their 2007 study, De Neys and Schaeken investigated the effect of load on the production of scalar implicatures. Their goal was to provide further evidence to the discussion whether scalar implicatures are generated automatically (neo-Gricean view, Levinson, 2000) or whether they are effortful (Relevance Theory, Sperber & Wilson, 1986/1995). In their study, they used a dual-task paradigm. Before completing a sentence verification task with scalar implicatures, participants, who were first year college students, had to memorize a complex dot-pattern. After every sentence verification, they had to recall the pattern and replicate it on an empty matrix. This secondary memorization task burdened the executive cognitive functions. If scalar inferences are indeed drawn automatically, then burdening the executive functions would not have an effect. Results of this study however, showed that participants did make less scalar inferences. This is direct evidence for Relevance theory.

In this experiment, we will look at both the effects of response measure and load on the production of scalar implicatures. We will use the same inference paradigm as used in Van Tiel et al. (submitted). Because of our secondary task and the presence of already two conditions in our experiment, we do not want to overcomplicate the design even further. Therefore, we use only a representative selection of the items used by Van Tiel et al. We will test some quantifiers and some gradable adjectives.

Our expectations with regard to different scales are, in accordance with Van Tiel et al. (submitted), that the quantifiers will lead to more pragmatic responses than the gradable adjectives. This partly because of the nature of the different scales, but also because the quantifiers we used are closed scales while the adjectives are all open scales; Van Tiel et al. already showed that closed scales lead to more pragmatic answers. We will create two groups of participants. One will have to judge scalar inferences in a classic binary fashion, the other group will have to judge them on a scale. We expect the scalar response type (i.e., the Likert-scale) to increase the rate of scalar implicatures. More specifically, we hypothesize less extreme average response patterns and more responses on the middle of the scale when participants use a scale to respond. Quantifiers elicited already up to 90% pragmatic answers in Van Tiel et al., which we assume is close to a ceiling effect. Therefore, we assume the effect of response measure will be greater for the gradable adjectives, which still have some space for improvement. Therefore, we expect an interaction between response measure and item type.

Additionally and congruent with De Neys and Schaeken (2007), we also manipulate cognitive load by using a complex dot pattern memorization task. This has proven to be an adequate way to burden participants' executive cognitive functions. We expect participants to give fewer pragmatic answers when they are highly burdened with a secondary task, and more when it is a light burdening. More concrete, this means that especially for the quantifiers, a decrease in pragmatic answers can be expected when a cognitive load is present. For the gradable adjectives, we already expect fairly logical answers. Therefore, we do not expect them to decrease as much as is the case for the quantifiers. We expect the participants without a secondary load task to have the highest rates

of scalar implicatures, for both the quantifiers and the gradable adjectives. Finally, we also expect there to be interactions between item and load, and item and response measure. We believe that these interactions will point to the possibility that not only do different scales elicit different levels of pragmatic responses, but that there are also different mechanisms behind the production of the different scalar implicatures. The different load conditions and response measures might then interact with these different mechanisms.

5.2.1. Method

The participants in this study were first year psychology students of the University of Leuven, who participated in return for course credit. Participants were expected to have at least 75% of the control items correct (see below for an explanation of the control items). Therefore, out of 371 participants, 79 were excluded due to bad performance on these control items. We had a 3×2 design with three between-subjects working memory load conditions and two between-subjects response type conditions. The three working memory load conditions were no load, low load and high load. For the low load and high load conditions, we used a classic spatial storage task, more specific the dot memory task, which has repeatedly been shown to be an adequate method to burden cognitive resources (e.g., Bethell-Fox & Shepard, 1988; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). In this dot memory task, the participants were shown a three by three matrix with several dots in it. The dots were either in a one-piece pattern of three dots (low load condition) or a two- or three-piece pattern of four dots (high load condition) (see Figure 1).

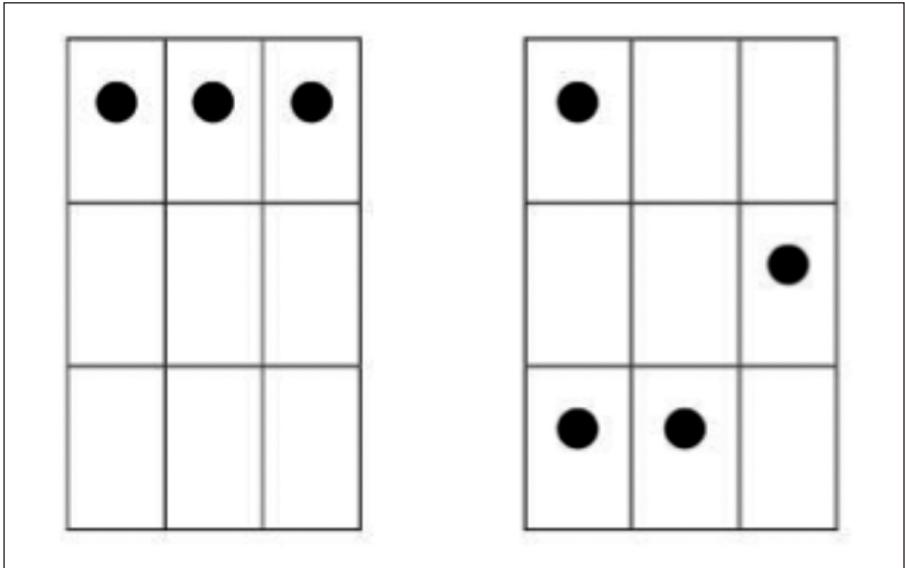


Figure 1. Examples of dot pattern matrices
(left: low load; right: high load).

This matrix was shown before the actual tested item, with the instruction to remember it. After the test item, participants were shown an empty matrix in which they had to replicate the dot pattern the way they remembered it. Work by De Neys, Schaeken, & d'Ydewalle (2005) has shown that the easy one-piece patterns only burden executive resources minimally. In the control no load condition, participants were not shown any dot matrix patterns. After the dot matrix was shown but before participants were asked to replicate the dot pattern in an empty matrix, a critical item was presented. Below is an example of such a critical item.

John says: The water in the bathtub is warm.

*Would you infer from this that, according to John,
the water in the bathtub is not hot?*

In each item, a person named John gives a statement which contains a scalar term. Afterwards, the participant is asked whether he or she can infer from the given statement that, according to John, a statement containing a stronger scalar term is not valid. In other words, the weaker term in the initial statement is replaced with the second stronger term and participants have to judge whether John would still agree with the altered statement. We used several different pairs of scalar terms: one pair of an existential quantifier (some/all), one pair of an epistemic modal (maybe/be certain) and five pairs of gradable adjectives. These scalar terms are a subset of the ones used in the study by Van Tiel et al. (submitted). We did not use all of their gradable adjective pairs, but only half and we attempted to take representative ones that would vary enough in the number of pragmatic responses they elicited. For each pair of scalar terms, we constructed five different experimental items. All together, this led to 35 experimental items in total. We also included 19 control items, of which 9 were clearly valid and 10 were clearly invalid. These control items were presented in a similar fashion as the test items. All the test items and filler items were randomized. Below are two examples of a clearly valid and a clearly invalid control item.

The movie was bad = The movie was not good.

The singer is tall ≠ The singer is not blond.

There were two different response measure conditions. Half of the participants were instructed to give a binary answer. They could either agree or disagree with the second statement. The other half of the participants had to indicate their answer on a 5-point Likert-scale. The scale ranged from completely disagree to completely agree.

5.2.2. Results

We converted all the binary responses 0 and 1 scores to 1 and 5 scores so that they would be comparable to the scalar responses. Therefore, for both the binary and the scalar responses, a response of 5 is a pragmatic interpretation of a scalar word and a response of 1 is a logical interpretation of a scalar word. Subjects performed as expected on the control items with the binary responses as well as with the scalar responses. Subjects in the binary response group had an average of 4.83 ($SD = .32$) for the valid items and 1.16 ($SD = .29$) for the invalid items. The average responses in the scalar group were 4.63 ($SD = .42$) for the valid items and 1.33 ($SD = .32$) for the invalid items, which are both significantly different from the binary group (valid items: $t(290) = 4.63, p < .001$; invalid items: $t(290) = 4.83, p < .001$). This difference can be expected as scalar answers are naturally less extreme than binary answers.

We ran a repeated measures design with the within-subjects variable item and two between-subjects variables, namely load (three levels: no load, low load and high load) and response measure (two levels: binary and scalar). We found a significant main effect of item ($F(7, 2002) = 652.06, p < .001$), a main effect of response measure ($F(1, 286) = 41.18, p < .001$) and a significant interaction between these two ($F(7, 2002) = 53.47, p < .001$). The main effect of load was only marginally significant

($F(2, 286) = 2.98, p = .05$) and there was no interaction with item.

The main effect of item can be explained by the extreme answer patterns for the control items, the existential quantifier and the epistemic modal, and the less extreme patterns for the gradable adjectives (see Figure 2). The main effect of response measure is clearly visible in Figure 2. The average for the participants in the binary condition is more pragmatic than the average of the participants in the scalar group. This is, as already mentioned, due to the nature of the different response measures. To explore the interaction between item and response measure more closely, we calculated *t*-tests for each item between the two different response types. All pairs were significantly different from each other, with all *t*-values between 3.59 and 9.44 and *p*-values all below .001. However, Figure 2 demonstrates that these differences are a bit smaller for the control items and the quantifiers. The control items receive extreme answers for both the valid and the invalid items, with little difference between the two response measure conditions. For the critical items, it seems that the existential quantifier and the epistemic modal are answered in a pragmatic way, for both the binary and the scalar responses. The adjectives however, are answered more logical in the binary response condition but a little more pragmatic when a scalar response is possible.

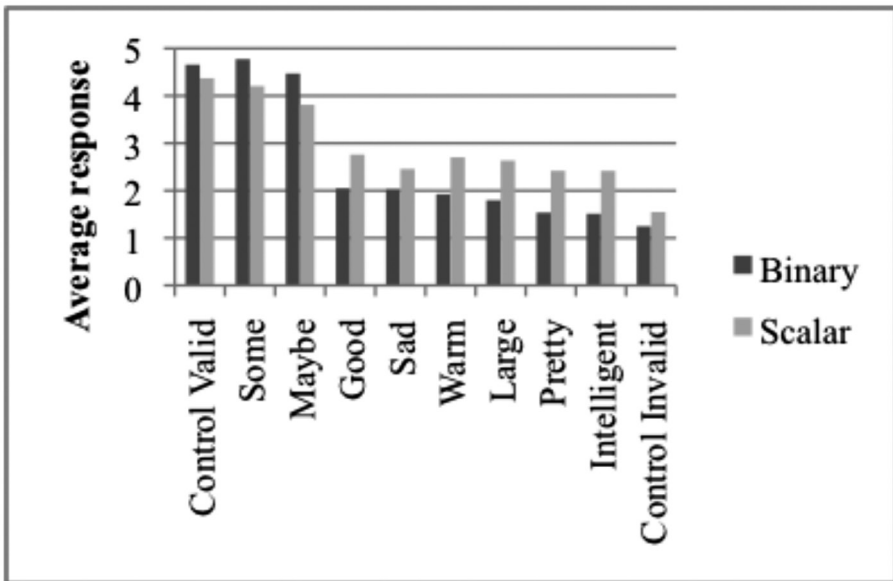


Figure 2. Average pragmatic responses for the different items, separated by response measure.

In Figure 3, we compare the frequencies of binary and scalar responses in the different item types. In this figure, we can see how the answers are divided over the different scale values. From our analysis we already know that participants choose fairly one-sided for all the different types with the binary option. When participants can answer on a scale, quantifiers are still answered very pragmatically. For the gradable adjectives however, the middle options of the scale become much more popular:

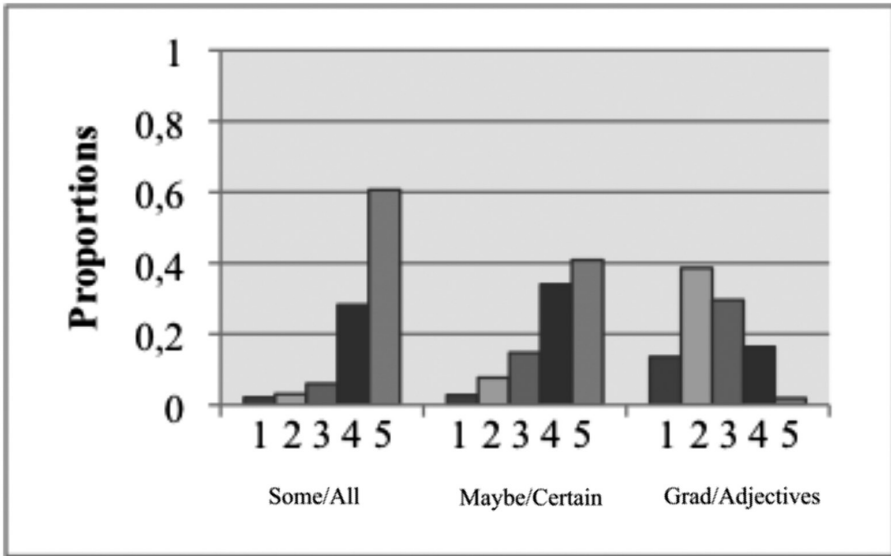


Figure 3. Proportions of scale values for the different item types.

5.2.3. Discussion

In this study we examined three different hypotheses. First, we expected different types of scalar inferences to evoke different rates of scalar implicatures. Therefore, we tested different types of scalar implicatures: two quantifiers and several gradable adjectives. We predicted that the quantifiers would lead to more scalar inferences than the gradable adjectives. Secondly, we predicted that response measure would play a role. More specifically, we expected that more scalar inferences would be produced when participants have to respond on a scale instead of giving a binary response. Thirdly, we predicted that the level of scalar inferences could be diminished when the cognitive resources of the participants are burdened with a secondary task.

Our results showed a main effect of item type. As we expected, the quantifiers were interpreted pragmatically more often than the gradable adjectives. Both the existential quantifier (some/all) and the epistemic modal (maybe/be certain) were judged extremely pragmatic. The gradable adjectives, however, were all answered logically. Additionally, there was a significant difference between the two response measures and there was an interaction of response measure with item type. In Figure 2, we saw that all items were judged less extreme when a scale was offered instead of the binary response option. However, the difference appears larger for the gradable adjectives than for the quantifiers. We believe this is partly because of a ceiling effect for the quantifiers. As we expected, parallel to Van Tiel et al. (submitted), the average response for the quantifiers with the binary response option was close to the maximum score of five. In Figure 3, it is illustrated that the binary pragmatic responses for the quantifiers are mostly divided over four and five, which is an extreme pragmatic response pattern. For the gradable adjectives however, the binary logical responses are much more distributed over all the different scale options. The three middle options on the scale take up more than 80% of the responses. So an extremely pragmatic answer on the gradable adjective scalar inferences seems very undesirable and it will not be used when it is not absolutely necessary.

The lack of a clear main effect of load was surprising. We will discuss several explanations for this in the general discussion.

5.3. Experiment 2

One of our concerns with the results of the first experiment, were the high rates of pragmatic responses for the quantifiers. Especially compared to other studies that investigate the use of scalar implicatures in adults, we see that the adults in our study score extremely pragmatic. For example in an experiment by Bott and Noveck (2004) adults were pragmatic in 61 percent of the time. The extreme answering patterns that we found make it very hard to interpret the results. We therefore believe it would be interesting to conduct a similar experiment in which the levels of pragmatic answers are lower. The way we will do this is by testing children.

It seems to be the case that children have more difficulty in grasping scalar implicatures and that they interpret implicatures differently than adults. More specifically, they interpret implicatures in a logical rather than a pragmatic way (Noveck, 2001). For this claim, we find evidence in different developmental studies. Smith (1980) shows that 4- to 7-year-olds treat the term *some* as compatible with *possible all*. They seem to interpret *some* in the logical way. Braine and Romain (1981) also show that 7- to 9-year-old children favor a more logical interpretation instead of a pragmatic interpretation. Noveck assumed that the explicit logical element of communication develops first in life and that the pragmatic element of communication develops later on. Following up on Smith (1980), Noveck compared 8-year-olds, 10-year-olds and adults on their capability of deriving the implicit meaning of *some*, by asking them to judge sentences on their truth-value. The Truth Value Judgment task (TVJT) they used, involved statements with the words *some* and *all*. In each condition, three kinds of statements were used: absurd ones, ap-

appropriate ones and underinformative ones. An example of the last-mentioned is 'Some giraffes have long necks'. The participants were asked to say whether or not they agreed with these statements. Noveck's main finding is that 8- and 10-year-old French children treated certain (some) as compatible with *tous* (*all*) much more than adults. Their answers were more logical. Furthermore, the 8-year-olds agreed more with the logical interpretations than the 10-year-olds. There seems to be a developmental trend going from 8- to 10-year-olds to adults in the derivation of scalar implicatures. This trend could be a consequence of the processing cost that is needed to make a pragmatic inference. Children have fewer cognitive resources available than adults, which makes them produce fewer pragmatic answers than adults (Noveck, 2001).

In the second experiment we will examine the processing of scalar implicatures by 5-year-old pre-schoolers. We will compare their responses on four different scales of implicatures, i.e. <some, all>, <maybe, certain>, <warm, hot>, and <good, excellent>. In line with previous research, we expect that quantifiers will elicit more pragmatic answers than other scales. We deliberately reduce the number of items. The reason for this is that we are testing very young children and their concentration spans are limited. The children will be divided into a binary and scalar response group. We expect that the Likert-scale will elicit more pragmatic answers. We will also have a look at the scores of the children on their Toetertest and whether this has any correlation with their capacity to make pragmatic inferences. The Toetertest is a standardized test that is used in Belgian pre-schools to test whether a 5-year-old child had acquired a number of skills that he will need in first grade. The test determines the extent to which a child has acquired the preparatory skills that he will need to learn to read, write and calculate. It focuses on

several sub-categories. The test will check for example whether a child can count, whether it can see the differences between two images or whether it understands words like 'highest' or 'less than'. Because it is very hard to implement a secondary task with children and children are already logical enough without implementing a load condition, we decided to replace the load condition with the results on the Toetertest. Because of the pragmatic-developmental trend, we expect lower scores on the Toetertest when there are more logic answers (and vice versa). We can assume that when there are more logic answers, the implicature demands more cognitive processing to answer pragmatic. We also assume that children who score lower on the Toetertest have less developed cognitive capacities that they need to reason pragmatically and so they will answer more logical.

5.3.1. Method

The sample consisted of 37 preschoolers (20 boys and 17 girls). The children were between the ages of 63 and 72 months ($M=60.73$, $SD=3.29$).

The stimuli were presented in short movie clips with puppets. We used this presentation method to make the stimuli more attractive for the preschoolers. In the clips one puppet uttered a sentence containing a scalar implicature, then a second puppet would rephrase the sentence with a denial of the stronger word on the scale. The children had to indicate whether the sentence said by the second puppet was correct. Below is an example of such an item.

Puppet 1: Some books in the school have an image.

Puppet 2: So not all books in the school have an image.

We used four sets of scalar terms: some vs. all; maybe vs. certain; warm vs. hot; good vs. excellent. Each set was presented six times, each time with a different context. We also included 12 control items which were either valid or invalid and we included 12 filler items. We decided to use a high number of filler items because it has been shown that a higher number of filler items makes participants less pragmatic and we wanted to make sure that our participants were not too pragmatic.

The children were divided into two groups. One group had to answer binary while the other had to answer on a 3-point Likert-scale, congruent to Katsos and Bishop (2011). The scale was made more attractive for children and was implemented in the stories told in the movie clips

5.3.2. Results

The binary responses 0 and 1 were converted into 1 and 3 so that we could compare them to the scalar responses. The critical items were converted so that a score of 1 would be a logical response and a score of 3 a pragmatic response.

We ran a repeated measures analysis with the within-subjects variable item and between-subjects variable response measure. There is a main effect of items ($F(3, 87) = 13.44; p < .001$), some items provoke more pragmatic responses than other items. There is no significant effect of response measure ($F(1, 29) = 1.92; p = .18$), so there is no difference between binary and scalar responses. Furthermore, there is no significant interaction effect between items and response measure ($F(3, 87) = 1.53; p = .21$). In Figure 4 we see a downward trend for the different items. Quantifiers provoke more pragmatic responses than the gradable adjectives.

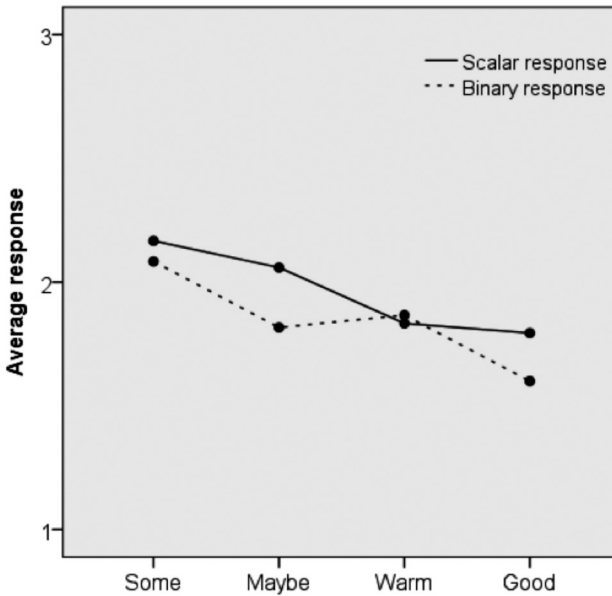


Figure 4. Average responses on the different items for the two separate tasks.

Paired sample *t*-tests were conducted to compare the answers on the different scalar terms. We compared six pairs of items of which five were significantly different from each other (Table 1). This explains the main effect of item.

To be sure that the responses on the items of each scalar term are equally distributed over the three response possibilities of the Likert-scale, we conducted a Chi-Square Test. The results show that the distribution of the responses on the Likert-scale differs across the items ($\chi^2(6, N = 37) = 15.91, p = 0.01$). When we take a closer look at the pairs of scalar terms, we see that only <some, all> versus <maybe, certain> ($\chi^2(2, N = 37) = 1.06, p = .59$) and <warm, hot> versus <good, excellent> ($\chi^2(2, N = 37) = 0.29, p = .87$) show the same distribution.

Table 2. Paired sample *t*-tests between all item types.

Pair of items	<i>t</i> -test	<i>p</i> -value
Some-Maybe	2.23	.033
Some-Warm	3.15	.004
Some-Good	6.69	.000
Maybe-Warm	1.32	.20
Maybe-Good	4.11	.000
Warm-Good	3.02	.005

Not only do some items elicit more or fewer pragmatic answers, but also the way of using the scale is different for most of the items. The distribution of the responses across the scale takes different shapes depending on the item.

When the children had to answer binary, their answers were fairly equally divided across the two answer possibilities (see Figure 5).

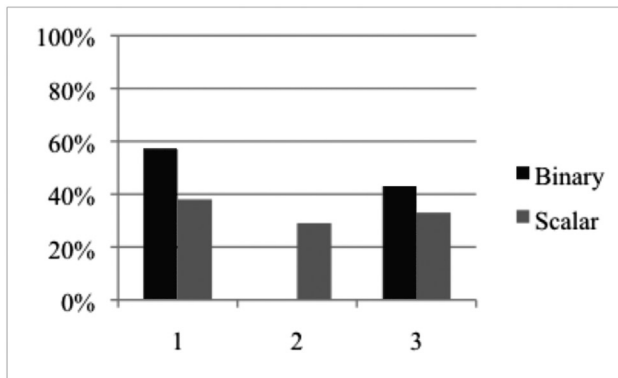


Figure 5. Comparison of the proportionate distribution of the responses on both tasks.

When we look at the correlations between the critical items and the scores on the Toetertest, we see that there is only a significant correlation for the critical item <warm, hot> ($r = .44; p = .01$). The children that scored higher on the Toetertest, tended to interpret this particular scalar implicature more logically. The scalar implicature <good, excellent> also provoked more logic answers. However, in this case we don't see a significant correlation with the Toetertest. There is a significant correlation between the responses of the scalar term <warm, hot> and the responses of the scalar term <good, excellent> ($r = .48; p = .01$). These implicatures seem to be 'harder' for the children, meaning that their answers are more logic on these items.

5.3.3. Discussion

In Experiment 2 we wanted to examine whether 5-year-old children process all scalar implicatures the same way. Therefore, we compared their responses on four different scales of implicatures, i.e. <some, all>, <maybe, certain>, <warm, hot>, and <good, excellent>. We expected, in line with previous research and Experiment 1, that children would process certain scales in a more pragmatic way and other scales more logically.

We were also interested in the effect of response measure. We used different response measures, a binary scale and a 3-point Likert-scale, as answer possibilities. We expected that the different response measures had an effect on pragmatic reasoning, whereby the Likert-scale would elicit more pragmatic answers.

Lastly, we were interested in the scores of the children on the Toetertest. We expected that lower scores on the Toetertest would be

an indication of less developed cognitive capacities. Therefore, these children would answer in a more logical fashion.

For our first hypothesis, the results were as we expected. Each type of scalar inference seems to be processed in a different way. We found that some scalar implicatures elicit more pragmatic interpretations than other scalar implicatures. It seemed that some scalar inferences require more cognitive effort to be decoded in a pragmatic way. The results contribute to the evidence against the uniformity assumption and are in line with Van Tiel et al. (submitted) and Experiment 1.

For our second hypothesis about response measure, the results are in contrast with our expectations and Experiment 1. There was no main effect of response measure. When the children had to answer binary, their answers were equally divided across the two answer possibilities. It is possible that one half of the participants reasoned logically and the other half reasoned pragmatically. Another possibility is that they randomly guessed. There is no way of knowing whether their responses were deliberate. Some children were consistent in their guesses, but others switched between pragmatic and logical responses. When the children had the possibility to answer on a Likert-scale, their responses were again equally divided across the three answer possibilities. This indicates that the children chose their answers by chance. The use of a Likert-scale should show a more nuanced response pattern and give more information than binary responses. It is remarkable that we did not find this result in this experiment. Our participants did not seem to produce more pragmatic answers when confronted with a Likert-scale, while other studies indicated that 5-year-olds were capable doing so. Our experimental design may be the cause of this phenomenon, which will be discussed later.

We could not confirm our last hypothesis. We looked at the correlations of the children's responses and their scores on the Toetertest. We only found a correlation with the scale <warm, hot>. The first experiment did not show evidence for the role of cognitive burdening either. Maybe the pragmatic-developmental trend is independent of other cognitive capacities. It is possible that the skills that are tested in the Toetertest are irrelevant for the pragmatic interpretation of scalar implicatures. The Toetertest consists of 11 different subtests and not all of these tests have any relevance to pragmatic reasoning. It is possible that the results of the Toetertest lie too far apart from the specific reasoning skills and vocabulary knowledge that is necessary to interpret scalar implicatures. Another factor that can play a role are the items that we used. Maybe our items were too easy for the children, even for the children who had low scores on their Toetertest. Or maybe it was the other way around, perhaps the items were too hard for the children. But we did find a correlation with the scalar term <warm, hot>. It is understandable that this item is the easiest critical item. We can assume that children are more familiar with concepts like 'warm' and 'hot'. Children are taught these concepts very early in life, because they are essential for one's own safety. A concept like 'excellent' may be not so well-known, making this item too difficult. Concepts like 'maybe' are more abstract which also makes them harder to learn or understand. However it seems unlikely that this is the case when we take into account the overall high scores on the control items. It might also be possible that the children did not understand properly what was expected of them and how they had to judge the statements. But again, the high performance on the control items indicates that they did understand the purpose of the experiment and the meaning of the scalar terms.

5.4. General discussion

These results clearly show that merely giving binary response options is inadequate for examining scalar inferences with adults. For children the results are more complicated. Using the scale provides a more nuanced image. Even though we could not replicate these findings for children, earlier research did find these effects and we believe it is due to the nature of our task that we could not replicate them. For future research it might be interesting to conduct these tests with a within-subjects design. This way we could see if the response pattern for each individual participant would change along with the type of answer possibility. We could investigate if there are clear response patterns or if the responses are still randomly chosen and children have no stable response strategy.

Either way, some general shortcomings in testing children this young must be addressed. We cannot be certain that the children were attentive the whole time. The attention span of children at the age of five is relatively small. We did our best to keep the test as short as possible. But even then it is hard to know how concentrated each child was. A solution to this problem would be to include more children in our experiment to give us more statistical power. In keeping the experiment as short as possible, we also included a relatively small number of items. A larger sample would be better, but we would again be confronted with the issue of concentration. Therefore, we stand by our original decision to use a short experimental design. The shortness of the experiment also induced a lot of variation, leading children having to change over different scalars all the time. Dieussaert, Verkerk, Gillard, & Schaeken (2011) found that a task was harder for adults when there was more variation between the different item types. This variation causes an extra load for

our working memory, having fewer cognitive resources left for pragmatic reasoning.

Our results show, even more than Van Tiel et al. (submitted), that there is a large difference between quantifiers and gradable adjectives. Not only is one group more pragmatic in their ratings and the other more logical, the variability between answers in each group shows that this difference is more complicated than originally expected. It is clear how the assumption of uniformity between scalar inferences is not correct.

While the lack of a clear correlation between the Toetertest and performance is easily explainable, the lack of an effect of load is more surprising. Several studies in the past have shown this effect yet we are not able to replicate it. We offer two possible explanations. One possible explanation could be found in Dieussaert et al. (2011). They conducted a study comparable to the one done by De Neys and Schaeken (2007). In their study, they tested a group of senior high school students. Similar to De Neys and Schaeken, the participants got a sentence verification task. While rating sentences like 'Some eels are fish' as correct or incorrect, they also had to memorize the complex dot patterns. Furthermore, the participants also completed a Dutch version of an Operation Span Task (OSPAN; La Pointe & Engle, 1990), which was adapted for group testing and computerized (GOSPAN; De Neys, d'Ydewalle, Schaeken, & Vos, 2002). This task served as a working memory capacity measure. Their conclusions complemented De Neys and Schaeken. First of all, they replicated the finding that a cognitive burdening decreases the rate of scalar inferences. Secondly and more importantly, they also found that this decrease is dependent on working memory capacity. People with high and low working memory capacity were less influenced by the cognitive load

than people in the middle memory capacity group. Therefore, it might be interesting to see whether performance in our experiment interacts with some general cognitive capacity measure. From a different experiment that our adult participants contributed to, we also had access to the performance scores of our participants on a Raven's Progressive Matrices Task (Raven, 1936). However, additional analyses did not reveal a pattern similar to the one in Dieussaert et al. (2011). Also, our participant group consisted of all first year university students. We can roughly derive from this information that this particular sample of participants most probably had high cognitive capabilities. They would probably all belong to the high performing group. Consequently, it is not that much of a surprise that no effect of load was found. However, the participants from the De Neys and Schaeken (2007) study were also first year psychology students from the same university and they did find an effect of load.

A different variable responsible for the lack of a load effect in this study may lay in the presentation of our stimuli. In light of our results, we wonder whether the classic verification paradigm might not have been better suited for our experiment than the inference paradigm that was suggested in Geurts and Pouscoulous (2009). As previously mentioned, performance seems to reach a ceiling effect for the quantifiers. Almost all the responses were extremely pragmatic, very close to a maximum score. We suspect that the difference between the binary responses and the scalar responses on quantifiers would be larger when the performance wasn't embedded in the maximum score. In De Neys and Schaeken, the average binary responses on the existential quantifier were lower than in our experiment. The only real difference between their study and our study being that they used a classic verification paradigm. It must be the case that the different presentation of stimuli in our study (similar to the

one in Van Tiel et al., submitted) makes the task much more prone to elicit pragmatic answers. This does explain the lack of an effect of load. When the primary task is too easy to begin with, a secondary task will not matter and no effect of load will appear. Therefore, we think it would be interesting to do a similar experiment with a classic verification task to explore whether the more difficult task does elicit a load effect.

From this study, we can confirm the conclusions from Van Tiel et al. (submitted). Not only are the rates at which scalar inferences elicit pragmatic answers dependent on task features, there is also a wide variability between the different scales that are used. Especially the <some, all> scale, which has been the most frequently used scale, seems widely unrepresentative and more an extreme case. We recommend future research to be careful when using a binary response method and recommend using a scale instead. Our results clearly show how participants' behavior changes drastically when they can answer on a scale. It is obvious how using a scale influences the rating of scales and gives a much more nuanced and realistic representation of scalar implicatures.

References

- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 12.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51, 437-457.
- Braine, M. D., & Romain, B. (1981). Development of comprehension of "or": Evidence for a sequence of competencies. *Journal of Experimental Child Psychology*, 31(1), 46-70.
- De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54, 128-133.
- De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and counterexample retrieval for causal conditionals. *Thinking & Reasoning*, 11, 123-150.
- De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, 42, 177-190.
- Dieussaert, K., Verkerk, S., Gillard, E., and Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology*, 64, 2352-2367.
- Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, 1, 1-38.
- Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.

- Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics & Pragmatics*, 2 (4), 1–34.
- Grice, P. (1989). *Studies in the Way of of Words*. Cambridge, MA: Harvard University Press.
- Horn, L. (1984). Toward a new taxonomy for pragmatic inference. In Schiffrin, D. (Ed.), *Meaning, Form and Use in Context: Linguistic Applications. Proceedings of GURT '84*. Washington D.C.: Georgetown University Press.
- Katsos, N., Andrés Roqueta, C., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119, 43–57.
- Katsos, N., & Bishop, D.V.M. (2011). Pragmatic tolerance: implications for the acquisition of informativeness and implicature. *Cognition*, 120, 67–81.
- Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness. In K. Franich, K.M. Iserman, L.L. Keil (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*. Somerville, MA: Cascadilla Press.
- La Pointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 1118–1133.
- Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, 130, 621–640.

- Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78, 165-188.
- Raven, J. C. (1936). *Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive*. MSc Thesis, University of London.
- Smith, C. L. (1980). Quantifiers and question answering in young children. *Journal of Experimental Child Psychology*, 30(2), 191-205.
- Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and cognition*. Oxford: Basil Blackwell.
- Van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B., (submitted). *Scalar diversity*.

6

Final Discussion

The focus of this dissertation was on scalar implicatures. We intended to explore scalar implicatures more in depth and also look at their development in children. We had two separate lines of research. In the first line, we focused on three features. We looked how task difficulty influenced the production of scalar implicatures, how the levels of pragmatic reasoning were influenced by the response measure that was used, and we explored reversed scalars. In the final chapter, we had a second line of studies, which focused on scalar diversity. In this chapter, we also looked at the effects of load on implicatures usage and the effects of response measure.

For our first line of studies, results were in accordance with what we expected. First of all, the response measure variable pretty much behaved as we expected it to. Like we saw in the literature (Katsos & Bishop, 2011; Katsos, Roqueta, Estevan, & Cummins, 2011; Katsos & Smith, 2010), the Likert-scale measure is so much more fine-tuned to investigate scalar implicatures. With the scalar response measure even children as young as 5 years old show performance levels comparable to adults and older children. Especially how the difference between children and adults on scalar implicatures can disappear completely like we showed in Chapter 2 is a convincing reason to start using these scalar response measures as the standard paradigm. Katsos and Smith (2010) describe the pragmatic skill behind implicatures as 'The ability to assess how much information should be communicated in a situation'. This exact ability might be the reason why children perform under the level of adults when responding binary. Perhaps children are fully aware of all the information that is available and that can be given, but they have difficulty with assessing which parts of this information is necessary and which is only optional. There are two different reasons why someone might reject

an underinformative utterance. First of all because it is underinformative and the person judges it could have been done better; secondly because the person makes the implicature in which case the utterance is incorrect. It is not easy to distinguish for which reason both adults and children reject an utterance. We need to know whether a participant notices that there is in fact a pragmatic violation in the first place. Subsequently, the participant needs to decide that this violation is grave enough to reject the utterance. Using a scale as a response measure in scalar implicature testing enables us to differentiate between these two steps. Both adults and children in our study indicated that they are competent enough to execute the first step. By using the middle of the scale, they can indicate that they are in fact aware of the pragmatic violation and that they possess sufficient pragmatic competence. The binary responses however, show us that adults and children differ in how they deal with this pragmatic violation. When given a binary response option, adults will reject the underinformative sentences while children will accept them more often. The pragmatic tolerance hypothesis is the explanation for this second step. Children are more tolerant towards this pragmatic violation than adults are. Using a binary measure is therefore not necessarily wrong. We can use it, as long as we are aware of what it measures, pragmatic tolerance and not pragmatic competence. Unfortunately, this distinction between tolerance and competence is not made in most studies. While most studies will intend to investigate competence, the results will unintentionally be influenced by the tolerance factor. This question of what exactly we claim to be testing is especially interesting for studies in which pragmatic reasoning is manipulated. For example, in the study by Papafragou & Musolino (2003) children were trained to detect underinformative utterances. More specifically, the experimenters enhanced the

children's awareness of the goals of the study. They found that this type of training led children to give more pragmatic responses. It is important to be careful in interpreting these results. If we interpret the results in light of the previously mentioned two steps, we believe it is more likely that by giving children training it is not as much their pragmatic competence that is influenced by the training, but their tolerance that is being sharpened. There are a few studies that influence the pragmatic tolerance in an even more direct fashion (Doran, Baker, McNabb, Larson, & Ward, 2009; Doran, Ward, Larson, McNabb, & Baker, 2012; Larson et al., 2009). These studies used a paradigm in which participants were asked to think how a fictional character named 'Literal Lucy' would judge the utterances. By doing this, they force participants to replace their personal pragmatic tolerance to that of a fictional character with a lower tolerance.

Whether or not the previously mentioned experiments altered the pragmatic tolerance of the pragmatic competence, is not something we can deduce directly from the data. Using a scalar response measure in this case could have given more fine-grained results. It would be interesting to see whether scalar responses are also influenced by the additional training. If the scalar responses are influenced by the training, it would indeed seem that training influences the pragmatic tolerance. If however the additional training influences the pragmatic competence and not the tolerance, we would not expect a difference in scalar responses between a trained and an untrained condition. A replication of the study conducted by Papafragou and Musolino (2003) for example, with a scalar response measure can prove this point. Especially the comparison between a condition in which children receive the additional instructions to make them aware of the goals of the study and a condition in which the children do not receive this additional instruction would be interesting.

An interesting study to look at within the framework of the two-step pragmatic reasoning process is the mouse-tracking study conducted by Tomlinson Jr, Bailey, and Bott (2013). Tomlinson, Bailey and Bott explained their results also as a two-step process. Participants start by interpreting a scalar inference in a logical way that is later enriched to a pragmatic interpretation. This change of heart is illustrated by the changing direction of the mouse. We can see the pragmatic enrichment as part of the pragmatic competence. Whether or not a participant changes direction is an indication of whether or not he is competent to make scalar implicatures. As a replication of this study we believe it would also be interesting to conduct a mouse-tracking study with scalar response measures. Perhaps it is possible to demonstrate the pragmatic tolerance with mouse-tracking as well. For starters we expect that participants that are confronted with a scalar implicature will start moving the mouse in the direction of the logical response on the scale but will end up clicking on the middle of the scale. If there is a process of pragmatic tolerance however, we can expect the participants to change the direction of the mouse twice. At first they will move towards the logical end of the scale. After they enrich their interpretation, the mouse will shift towards the pragmatic end of the scale and the pragmatic competence process is completed. At this point however, the pragmatic tolerance process takes place. The participant will be aware of the two conflicting interpretations and needs to decide how to cope with them. As a final movement, the participants will change the direction of the mouse towards the middle of the scale. This mouse track would illustrate both pragmatic competence and pragmatic tolerance.

Task difficulty proved to be a significant factor for pragmatic reasoning as well. We found that children were highly influenced by the

difficulty of the task. More difficult tasks led to less pragmatic reasoning. This is in accordance with the Context-driven account. If the pragmatic interpretation of a scalar implicature requires cognitive effort, then we would expect children, who have fewer cognitive resources available, to be less pragmatic than adults. This hypothesis was confirmed by our data. Children were less pragmatic than adults, especially on the more difficult tasks, Euler Circles and Immediate Inferences. The preschoolers we tested were not influenced by task difficulty. However, this might be due to the tasks we used. The way in which the tasks were made more or less difficult in the experiments we used in Chapter 2, was by altering the level of abstract reasoning that was required. For example the Euler Circles task that we used in Experiment 1 and 2 of Chapter 2 is a very abstract task and therefore fairly difficult for children. The Drawings task that we also used in these experiments on the other hand is much more heuristic and thus easier for children. This difficulty variable led to a large difference between the average responses of children on the scalar implicatures for the Euler Circles task compared to the Drawings task. While the average responses on the Euler Circles task were on the lower end of the scale, the responses on the Drawings task were much more pragmatic on the middle of the scale. This difference was not present for adults. For the preschoolers in Chapter 3 however, we had to modify the abstract task because it would be too hard for children this young. At the same time, we wanted to keep the task as similar as possible to the one we used previously so we could compare the results. Instead of having an abstract representation of blocks in Euler Circles, we used actual circles and actual blocks. Because of the real blocks, the circles became redundant. Children could answer the questions solely with the blocks and without the circles. We kept them anyway, to make the task even simpler and to keep

the similarity with the abstract task. It is possible however, that because of these alterations, the task became too easy and that there was no difference in difficulty with the drawings task left. As a result, we did not find an effect of task difficulty. Moreover, one can argue that the original Euler Circles task elicits logical thinking about scalar implicatures because the task in itself requires logical abstract thinking. By altering the task to a less abstract version, it is possible that these logical thinking skills are provoked less too. The heuristic nature of the Drawings task does not elicit logical thinking as much as it elicits pragmatic thinking. Again, this would explain why we did not find a main effect of task difficulty.

We did not find an effect of task difficulty for adults either in Chapter 2. We assume all the tasks were easy enough for adults. Even though the tasks varied in difficulty, if they are all still easy enough, adults will have enough spare cognitive resources available to think pragmatically. This pattern was shown in the performance on all the different tasks, it was very high for both the Euler Circles task, the Immediate Inferences as the Drawings task. In experiment 1 and 2 of Chapter 2, all the control items were near perfect, and all the scalar implicatures were answered pragmatically. For both the Euler Circles task and the Drawings task, performances on the scalar implicature were on the lower half of the scale. The scores between the different tasks were not significantly different from each other. If the adults are already pragmatic on the most difficult task then there was not much room to become even more pragmatic on the easier tasks. There are two options to look at the effects of task difficulty. Either we investigate subjects who are less pragmatic, for example children, and make a comparison to adults; or we make adults less pragmatic by adding a load variable. We executed the first option in Chapters 2 and 3, in which we tested both children and preschoolers. But the second

option would be interesting too, to test cognitively burdened adults with scalar response measures. Adults were already tested in several studies with a load factor and a binary response measure and in these studies effects of load were found (De Neys & Schaeken, 2007; Dieussaert, Verkerk, Gillard, & Schaeken, 2011). It would be interesting to look at the interaction between load and response measure. This interaction would be interesting because it could reveal how load influences the two steps of pragmatic reasoning, pragmatic competence and pragmatic tolerance, separately. It seems more plausible to us that load only affects pragmatic competence and not tolerance. Making a comparison between scalar responses and binary responses could give us more insight in this matter. In the last study of this dissertation, we included both load and response measure but we did not find an interaction between these two. However, these variables were not the primary focus of this study and we used a large set of different scalar terms. Our study showed that pragmatic reasoning differed greatly on these items and that the items interacted with the response. Therefore, we do not want to draw conclusions from this study about the effect of load on the two steps of pragmatic reasoning. It is very likely that the different items clouded our results too much. We believe a simpler study with only one Horn scale would be more appropriate. On top of that it would be interesting to see how these results interact with working memory capacity. Dieussaert et al. (2011) found in their study that load only affects participants with a smaller working memory capacity. If we look at this from the two steps of pragmatic reasoning point of view we can assume that load only affects pragmatic competence and moreover that it only affects the pragmatic competence of participants with low working memory capacity. If this is true, we can expect that we will find an three-way interaction between

response measure, load and working memory capacity. We expect that for the binary response measure the effect of load will be larger for the group with the lower working memory capacity. For the scalar response measure however, we would expect this difference to be smaller. As the scalar response measure only influences pragmatic tolerance and not pragmatic competence, than both groups should be able to acknowledge that there is a conflict between the two interpretations. This argumentation only works of course if pragmatic tolerance is not influenced by our working memory capacity. If tolerance is in fact influenced by it, then we only expect to find the interaction between load and working memory capacity, congruent to Dieussaert et al. (2011), but no three-way interaction with response measure.

It is not clear how the task of the violation tolerance experiment rates on task difficulty. If we follow the same argument we used earlier about easy versus difficult tasks for this experiment, it is possible that the task was too easy, not leaving any room for variability between items. We still see this variability for the *some* items, but not for the *all* items, like we expected. We found that for the *all* items, participants almost always used the top and the bottom of the scale and barely ever used the middle. It would be interesting to see what would happen if participants carried out this task under the burden of a secondary task. This burden would probably lead to fewer pragmatic responses for the *some* items, as we have already seen in the literature. But perhaps it would also influence the *all* items and change the pattern in the direction that we predicted with the violation tolerance theory.

The most interesting finding of this dissertation was definitely the occurrence of reversed scalars. We found that reversed scalars are not treated as the control items they are by children. In Experiment 1 and 2

of Chapter 2 we showed that children use the middle of the scale to rate these items, indicating some form of uncertainty or ambiguity about these items. It remains unclear why this particular item type has not been studied in previous research, as it seems to be the perfect control item for the normal scalar implicature. Especially the claim that scalar implicatures are formed because of the one-directional entailment of words on a Horn scales is being compromised. Our reversed scalars findings do not agree with this claim. If scalar implicatures are possible because the weaker term entails the stronger term and that it can therefore be used instead of the stronger term, then reversed scalars would not be possible. The stronger term does not entail the weaker term and using the stronger when the weaker is appropriate, is incorrect. Nevertheless, children handle these reversed scalars like ambiguous items and not like the control items they actually are. When responding binary, both adults and children acknowledge that these items are false. When they respond on a scale though, children especially treat these items differently and rate them as equally correct as incorrect. Adults are not as obvious as children in this pattern, but still they do show some of that tendency like children. They definitely do not treat them like the other control items. When we look at the preschoolers in Chapter 3, the trend becomes even clearer. For the preschoolers the trend is even visible for the binary responses. Almost half of the children accept these statements to be correct. Especially the results we found with the adults suggest that there is more at play in the interpretation of these scalar terms. They suggest that some general assumptions about scalar implicatures do not hold and that more factors must be at work. The results we found for the preschoolers, combined with the results we found for the older children and the adults, show a developmental trend. While it seems very

hard for preschoolers to distinguish these items from scalar implicatures, older children already feel they are incorrect at some level, which is expressed in the binary results. Adults lastly, are even better than children at classifying these items as incorrect. This developmental trend is a clear indication that the results from the violation tolerance experiment could have been completely different if it were tested with children. The fact that adults are capable of rejecting these items binary, while preschoolers are not, is a clear indication that there is some skill that adults have developed, or some knowledge that adults have acquired, that children do not have yet. We believe this might be the skill to differentiate between logical and pragmatic violations. This skill is probably a mix between pragmatic competence and pragmatic tolerance in the two-step hypothesis. The fact that adults and older children are able to reject them binary and preschoolers are not, suggests that this is a subtle skill. Perhaps the task we used to verify our hypothesis was too obtuse to pick up on this subtle skill. We assume that preschoolers might still have some serious issues with pragmatic competence, which leads them to not even reject the reversed scalars when responding binary. Older children will have mastered the necessary pragmatic competence, but still have some issues with the pragmatic tolerance, compared to adults. Adults finally, have mastered both skills, and while they do not reject scalar implicatures when responding on a scale, they do reject the reversed scalars.

We are fully aware of the gaping hole in our experiment. Our violation tolerance experiment should have been carried out with children. Unfortunately because of the time constraints in doctoral research, we did not have the time to carry out this study. We can therefore only speculate on what the results of that study would have revealed. We believe that our results from the first experiments clearly show a wide

discrepancy with the accepted theories and that it is a matter of exploring what causes this discrepancy. Even though the results of our experiment did not favor the violation tolerance hypothesis over the pragmatic tolerance hypothesis, we still stand behind it.

We want to reflect on how our experiments, especially the violation tolerance experiment, can be reconciled with the Context-driven account. The Context-driven account states that participants will produce a logical interpretation first, and later on, only when the context provokes it, the pragmatic interpretation will be produced and used. The high levels of pragmatic answers with the children and adults we tested show that our tasks do have enough context in them for the pragmatic interpretation to be triggered. We wonder though how much of that context information is available in the violation tolerance task. On the one hand, one could argue that there is enough because the participants give pragmatic responses on the scalar implicatures. On the other hand, our task with the black and white dots is extremely abstract. It is a very unnatural task and one can question its ecological validity. There is only one way to interpret the task, there is no additional context given and especially the *all* and *none* terms have only one very straightforward interpretation possible. If there is no context given what so ever, how can we speak of a context-driven response? Even more so, with no context given and overall extremely pragmatic responses, we can question whether those pragmatic responses that were given were just default, instead of the result of a costly cognitive reasoning process. This brings us back to the matter of cancellability. An implicature can be cancelled when the context asks for it. In this experiment, there is however no context and cancelling the implicature is not very easy, with the task being reduced to a semantic matching of a sentence and a picture. Several authors have

argued though that cancellability is a necessary feature of conversational implicatures (Blome-Tillmann, 2008; Borge, 2009; Grice, 1978). For a task like this one, we wonder if we can even still talk about pragmatic versus logical reasoning. This distinction suggests that there are multiple ways to interpret the sentences, which there are not in this case. The only possible way to interpret the *all* and *none* sentences is the semantic way. There seems to be not much room for the reversed scalars to manifest like there was in the tasks of the previous experiments. Imagine that we tested children with the violation tolerance task and their performance was as we expected, with a slow inclination for the *all* sentences. How should we interpret this? Would this be pragmatic or logical reasoning? Again we can say this is semantic reasoning that has gone wrong. Perhaps children do not fully understand the meaning of the word *all*. This particular idea will be discussed below, but for now, let's assume that they do understand it. Children have repeatedly shown that they understand these words correctly, which is shown in all the other control items we presented them with, and it has been shown in the literature as well. More likely would be our violation tolerance hypothesis. Children think in a more heuristic way. They know all the meanings of the words and all the rules they are supposed to be following. They are merely less scrupulous about those rules, independent of whether those rules are pragmatic or semantic. This would mean children are capable of a certain level of pragmatic thinking. They do not thoughtlessly follow the rules; they play with them and mold them in a way they see fit. This would mean that children play with these rules and think pragmatically more than adults do, while adults think more along the lines of some well-formed rules. The Context-driven account predicts logical reasoning to happen first, and therefore to develop first in children too. The idea that children play

with the rules does not support the idea of the Context-driven account that we start our development by reasoning logically.

As previously mentioned, we assume that children know the correct meaning of words like *some* and *all*. We assume that they know the meaning, but that they are not very strict about semantic rules. We can link this to the two-step pragmatic reasoning process that we mentioned earlier. First of all, participants need to have a sensitiveness of informativeness; they have to notice that an utterance is underinformative and that there are two ways to interpret it, which was called pragmatic competence. Secondly, they have to decide that this violation is grave enough to punish it, pragmatic tolerance. We previously stated that we assume that the problem for children with scalar implicatures lies in the pragmatic tolerance, which was shown by the scalar responses. Children know the semantic meaning of the words, they are pragmatically competent, but their pragmatic tolerance is different from the pragmatic tolerance of adults. But what if there is a problem with the pragmatic competence as well? It is possible that the words are not fully understood yet. Barner, Brooks, and Bale (2011) already showed that preschoolers have a hard time coming up with alternative meanings to words like *some* and that the meanings of words like these is clearly not as well defined as they are for adults. Perhaps children have a fuzzy concept of what these words mean exactly. Obviously they have some general idea of what they mean, which enables them to give correct responses in control items that do not hold any conflict to them. When it comes to scalar implicatures and reversed scalars however, the semantic distance between these words is so small that it becomes harder to differentiate between them. In this case, children's responses on the middle of the scale might not mean that they are aware of the interpretational conflict like it does for adults

but instead mean that they are not really sure, it represents the fuzziness. The results of a carrying out the violation tolerance task with children would not only give us insights about the violation tolerance hypothesis, but also clarify this issue. If reversed scalars were the result of fuzziness, then we would not predict a linear trend for the some items like we do in the violation tolerance hypothesis. Instead, we would expect them to rate all the different ratios (except for 0/8 and 8/8) somewhere in the middle of the scale, because the term *all* would be fuzzy independently of the ratio of dots.

For our final study, results were only partially in line with our expectations. Like we expected, response measure showed, yet again, that the interpretation of scalar implicatures is not a black or white story. There is a clear middle ground and offering a scalar response option is a way to reveal this compromise. Secondly, people treat the different scales differently. Not all scalar implicatures are processed the same way. Especially for gradable adjectives it seems hard to come to a pragmatic interpretation, which is easier for the existential quantifier and epistemic modal. Like expected, the *some* and *all* scale is not a representative example of scalar implicatures. Yet this exact scale seems to be used in research over and over again. Do people assume there is uniformity between all scales? We acknowledge that to do research, you need to pick out an example to use and certain easiness arises when everybody uses the same example. We used that very same example in all our studies too, and we made frequent comparisons to other studies on scalar implicatures. We realize that these comparisons are only possible because everybody uses this same flagship example. But there are certain risks associated with using only one example. We all make claims about scalar implicatures as a whole, based on this one example. By doing this, we

unwittingly have promoted this example to the prototypical example of scalar implicatures. Our study however shows that there is in fact large variability in how the different examples are processed. It might be possible that the *some* example is in fact a prototypical example of how scalar implicatures work. It is however also possible that *some* is as prototypical to scalar implicatures as a whale is to mammals. We have to remember that we only tested a small sample of scales in our study. In the study by van Tiel, van Miltenburg, Zevakhina, & Geurts (2013) a more extended set of 43 scales was tested. Thirty-eight of these scales however were adjectives and verbs and only 2 were quantifiers. This is a fairly biased set and probably not a representative set of scales that are used in our everyday language. Either way, a certain caution is clearly advised when interpreting results. Although several studies have already been conducted on scalar diversity (Beltrama & Xiang, 2012; Doran et al., 2009; van Tiel, van Miltenburg, Zevakhina, & Geurts, 2013; Zevakhina, 2012), more research is the only way to bring more clarity. As far as we know, none of the studies on scalar diversity have looked at more than 50 scales at a time. With the wide extend of our vocabulary this number seems pit-eous. Perhaps it is possible in the future to conduct a study similar to the one by van Tiel et al. (2013) with a much more extensive set of scales and find an actual prototypical example of a scalar implicature.

In light of this typicality discussion, we want to point out that one can also question whether scalar implicatures are a good prototype of pragmatic reasoning. In our dissertation we have made several claims about pragmatic competence, all based on scalar implicature research, but it is possible that scalar implicature reasoning is not a representative example of pragmatic reasoning. We already mentioned in our introduction that research on Autism Spectrum Disorder (ASD) showed that

there are differences between different subtypes of ASD when it comes to pragmatic inferences, but that these differences do not exist for other types of pragmatic reasoning (Baltaxe, 1977; De Villiers, Stainton, & Szatmari, 2007; Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts, 2009; Surian, 1996). This shows that there are in fact differences between different types of pragmatic reasoning. We do not know whether all the claims that we have made about scalar implicatures can be generalized to other types. It would be interesting to experimentally investigate the claims we have made about pragmatic competence on other types of pragmatic reasoning.

Lastly, we did not find an effect of load. This is surprising as previous research does indicate that pragmatic reasoning is influenced by load (De Neys & Schaeken, 2007; Dieussaert et al., 2011; Marty, Chemla, & Spector, 2013; Marty & Chemla, 2013). Why we were not able to replicate this finding remains unclear to us. Even in previous unpublished work, we have not been able to replicate this effect of load. We found other studies that did not find an effect of working memory either (Banga, Heutinck, Berends, & Hendriks, 2009; Janssens, Fabry, & Schaeken, 2014). It makes us wonder how robust the load effect actually is. We did however find an effect of task difficulty in the experiments in Chapter 2. Even though we tested the effects of task difficulty in the those experiments and load in Chapter 5, we do believe we can make a direct comparison between these two and some considerations on this subject might be in order. We used task difficulty in our first experiments, but we believe it shares the same background as load. We believe that pragmatic reasoning required effort and sufficient cognitive resources should be available to support it. Both a load variable and task difficulty influence pragmatic reasoning because of this. If we add a secondary task, or a load

variable, we insert a large part of our available resources in that task, and not sufficient resources are available to reason pragmatically. If we use a very difficult primary task however, we also burden our cognitive resources a lot. A lot of energy goes into the basic requirements of the task and not enough resources are left for the pragmatic reasoning part. Therefore, the end results of a secondary task and a difficult task are the same, our pragmatic reasoning is impaired, and a direct comparison between the two should be possible. If we compare with our first few experiments, for which we only found an effect of task difficulty for children, but not for adults, this lack of a load effect is not surprising. But then why are children sensitive for this extra variable and why are adults not? According to the Context-driven account, making the implicature requires effort. But perhaps our load manipulation was not strong enough for adults, leaving them with more than enough cognitive resources to make the implicature. Children however have fewer resources to begin with, and perhaps the most difficult task was more demanding to them than the secondary task was to adults. The secondary task we used however has proven to be an adequate load factor for adults in studies before ours (De Neys & Schaeken, 2007; Dieussaert et al., 2011), even outside the research area of scalar implicatures (Bethell-Fox & Shepard, 1988; De Neys, 2006; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; Verschueren, Schaeken, & d'Ydewalle, 2004). Why did it not work this time? Perhaps it has something to do with the participants that we used in our experiment. Dieussaert et al. (2011) showed that the effect of load only has an influence on participants with a small working memory capacity. The participants in our study however were all first year university students. We can make some general assumptions about first year university students, that they are more intelligent than the average

population. Perhaps they also have, on average, larger working memory capacities than the average population, which leads them to not being susceptible to a secondary load factor. It is also possible that making the implicature is 'demanding' in another way than remembering our dot patterns was. Perhaps it is based on another skill set than the one that is burdened in the secondary task that we use. Marty et al. (2013) found that the effects of load are different for the stimuli with *some* than they are for stimuli with numerals. While a dual-task paradigm has a diminishing effect on pragmatic reasoning about *some*, it has a stimulating effect on numerals. Of course one can question the place of numerals within the scalar implicature research all together (Geurts, 2006; Spector, 2013) but nevertheless it is clear that the 'demanding effect' of load on scalar implicatures is not a clear picture.

We also believe that the robustness of pragmatic reasoning of children plays a role in why we did not find a load effect. Multiple studies have shown how easy it is to influence pragmatic reasoning in children (Papafragou & Musolino, 2003; Papafragou & Tantalou, 2004; Pouscoulous, Noveck, Politzer, & Bastide, 2007). When children are so easily affected by task features, it is only natural that they are susceptible to task difficulty or load. For adults however, pragmatic reasoning is much more established and it is much less susceptible to outside factors like load or task difficulty. We suggest a theory that could explain this difference between children and adults and that could explain the lack of a load effect in adults, which originates from the intuitive feeling that we think according to the Default account. Is it possible that our brain, as it matures, evolves from a Context-driven account way of thinking towards a Default account way of thinking? Perhaps, when we are children, we start with learning the logical semantic meaning of a word and only later on do we learn to reason

more pragmatically. As we become older however, this pragmatic way of reasoning becomes our default way of thinking. Before we have matured to this pragmatic default, we are still susceptible to load or task difficulty. As soon as the pragmatic process does become default, we are not susceptible to them anymore. When we look at the study by De Neys and Schaeken (2007) that did find a load effect on pragmatic reasoning, we see that they tested first-year college students. Although no specific demographics are given in the article, we can assume an average age of 18-19 years old. Most studies consider this particular demographic to fall within the 'adults' group. But is it possible that when it comes to pragmatic reasoning, 18 year-olds are still evolving? Studies have shown that adult-level mature performance of working memory only began at the age of 19 years old (Luna, Garver, Urban, Lazar, & Sweeney, 2004; Luna & Sweeney, 2001). It is not unlikely that a delicate skill like pragmatic reasoning also takes a long time to develop completely and perhaps by the age of 50, we have evolved towards a Default account way of thinking. Other studies on reasoning processes in elderly have been conducted, but to our knowledge, there have been no studies on the production of scalar implicatures in late adulthood. A comparative study between children, young adults and older adults might hold prove to our theory. If there is a significant difference in the interpretation of scalar implicatures between young and older adults, this might point into the direction that the development of pragmatic reasoning is not finished at the age of 20. If this difference exists, it would also be interesting to replicate some of the experiments that are able to differentiate between the Default Theory and the Context-driven Theory with elderly participants. For example it could be interesting to look at reaction times, similar to Bott and Noveck (2004). Perhaps older adults do not need more time to pragmatically

interpret a scalar implicature. If they are in fact faster with the pragmatic interpretation than they are with a logical one, this could point into the direction of an evolutionary theory of pragmatic reasoning.

This dissertation has given us several interesting findings. First of all, we have showed that task difficulty and response measure can play a large role in the interpretation of scalar implicatures. We have showed that these task features can influence the levels of pragmatic reasoning immensely. Secondly, we have pointed out that, as a result of these task features, it is important to keep a close eye on what exactly is being tested or manipulated in a task. Pragmatic competence and pragmatic tolerance are two separate steps and it is necessary to make clear predictions on which of these two steps one intends to influence or investigate. Thirdly, we have given some strong evidence, in the form of reversed scalars, that put pressure on frequently used theories of scalar implicatures. These particular items have not been studied enough in the past and they may possibly hold the key to unraveling the process behind scalar implicatures. Finally, we have reconfirmed that there is in fact a large scalar diversity and that one must be careful in using the 'some' example as the prototypical example scalar implicatures.

References

- Baltaxe, C. A. (1977). Pragmatic deficits in the language of autistic adolescents. *Journal of Pediatric Psychology*, 2(4), 176–180.
- Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. *Linguistics in the Netherlands*, 26, 1.
- Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, 118(1), 84–93.
- Beltrama, A., & Xiang, M. (2012). Is good better than excellent? An experimental investigation on scalar implicatures and gradable adjectives. *Proceedings of Sinn und Bedeutung*, 17, 81–98.
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 12.
- Blome-Tillmann, M. (2008). Conversational implicature and the cancellability test. *Analysis*, 68(2), 156–160.
- Borge, S. (2009). Conversational implicatures and cancellability. *Acta Analytica*, 24(2), 149–154.
- Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3), 437–457.
- De Neys, W. (2006). Dual processing in reasoning two systems but one reasoner. *Psychological Science*, 17(5), 428–433.
- De Neys, W., & Schaeken, W. (2007). When People Are More Logical Under Cognitive Load. *Experimental Psychology*, 54(2), 128–133.

- De Villiers, J., Stainton, R. J., & Szatmari, P. (2007). Pragmatic abilities in autism spectrum disorder: A case study in philosophy and the empirical. *Midwest Studies in Philosophy*, 31(1), 292–317.
- Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: further evidence that scalar implicatures are effortful. *The Quarterly Journal of Experimental Psychology*, 64(12), 2352–2367.
- Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, 1(1), 211–248.
- Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. *Language*, 88(1), 124–154.
- Geurts, B. (2006). The meaning and use of a number word. *Non-Definiteness and Plurality*, 95, 311.
- Grice, P. (1978). Some further notes on logic and conversation. In P. Cole (Ed.), *Syntax and semantics, volume 9: Pragmatics*, (pp. 113–127). New York: Academic Press.
- Janssens, L., Fabry, I., Schaeken, W. (2014). 'Some' effects of age, task, task content and working memory on scalar implicature processing. *Psychologica Belgica*, 54 (4), 374-388.
- Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1), 67–81.
- Katsos, N., Roqueta, C. A., Estevan, R. A. C., & Cummins, C. (2011). Are children with Specific Language Impairment competent with the pragmatics and logic of quantification? *Cognition*, 119(1), 43–57.

- Katsos, N., & Smith, N. (2010). Pragmatic Tolerance or a speaker-comprehender asymmetry in the acquisition of informativeness? In K. Franich, K.M. Iserman, L.L. Keil (Eds.), *Proceedings of the 34th Annual Boston Conference in Language Development*. Somerville, MA: Cascadilla Press.
- Larson, M., Doran, R., McNabb, Y., Baker, R., Berends, M., Djalili, A., & Ward, G. (2009). Distinguishing the said from the implicated using a novel experimental paradigm. In U. Sauerland, & K. Yatsushiro (Eds.), *Semantics and pragmatics: from experiment to theory*, (pp. 74–93). Berlin: Palgrave MacMillan.
- Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A., & Sweeney, J. A. (2004). Maturation of Cognitive Processes From Late Childhood to Adulthood. *Child Development*, 75(5), 1357–1372.
- Luna, B., & Sweeney, J. A. (2001). Studies of Brain and Cognitive Maturation Through Childhood and Adolescence: A Strategy for Testing Neurodevelopmental Hypotheses. *Schizophrenia Bulletin*, 27(3), 443–455.
- Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, 133, 152–163.
- Marty, P.P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology*, 4.
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, 130(4), 621.
- Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3), 253–282.

- Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition*, 12(1), 71–82.
- Pijnacker, J., Hagoort, P., Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, 39(4), 607–618.
- Pouscoulous, N., Noveck, I.A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, 14(4), 347–375.
- Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, 7(5), 273–294.
- Surian, L. (1996). Are children with autism deaf to gricean maxims? *Cognitive Neuropsychiatry*, 1(1), 55–72.
- Tomlinson Jr, J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69(1), 18–35.
- Van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (submitted). *Scalar diversity*.
- Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2004). Everyday conditional reasoning with working memory preload. In *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 1399–1404). Austin, TX: Cognitive Science Society.
- Zevakhina, N. (2012). Strength and similarity of scalar alternatives. *Proceedings of Sinn und Bedeutung*, 16, 647–658.

