Children's Ability to Malinger Cognitive Deficits

by

Jesse Emil Elterman

M.A. (Psychology), Simon Fraser University, 2007
B.A. (Psychology), University of Victoria, 2005

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

Department of Psychology

Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Spring 2018

Approval

Name: **Jesse Emil Elterman**

Degree: **Doctor of Philosophy**

Title: ***Children's Ability to Malinger Cognitive Deficits***

Examining Committee: **Chair:** John McDonald, Ph.D.
Professor

Deborah Connolly, LL.B., Ph.D
Senior Supervisor
Professor

Kevin Douglas, LL.B., Ph.D.
Supervisor
Professor

Allen Thornton, Ph.D., R.Psych.
Supervisor
Professor

Charlotte Waddell, MD, FRCPC
Internal Examiner
Professor
Faculty of Health Sciences

Jennifer Batchelor, Ph.D.
External Examiner
Director Clinical Neuropsychology
Department of Psychology
Macquarie University

Date Defended/Approved: January 24, 2018

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Historically, psychologists have not assessed performance validity in child assessments where there is potential for secondary gain, although children's ability to malinger is a growing concern among psychologists. The present series of three simulation studies examined (1) children's ability to withhold their best effort on psychological testing, (2) whether performance validity tests (PVTs) can accurately detect withholding, and (3) whether inhibitory control explains individual differences in withholding. Participants were children in grades four and six, and performance was measured using the WISC-IV PSI subtests and the RAVLT. PVTs included the MSVT and TOMM. Children were instructed to either try their best (BE condition) or withhold (WE condition) based on instructions in a storybook read to children by their parent/guardian the night before the testing session. Study 1 used a repeated measures design to first assess best effort and then children were either given the BE or WE storybook. Both cohorts of children in the WE condition performed worse on the RAVLT and PSI than children in the BE condition, the PVTs obtained moderate accuracy, and inhibitory control was unrelated to withholding. Study 2 examined whether prior exposure to the testing materials was necessary for children to withhold and only included one testing session. Children in grade 4 scored lower on the PSI than their best effort comparison group, but otherwise scores on the performance tests were not statistically different between groups based on instructions given. There were significant differences between the groups on the MSVT, although the classification accuracy was only moderate. Study 3 examined whether children could withhold their best effort without being reminded of the instructions prior to the testing session. There were no differences on the performance tests or PVTs between those instructed to withhold and the best effort comparison group, although the variances on the MSVT were unequal, suggesting that the instructions had an effect, albeit subtle. Overall, the results show that fewer children can withhold their best effort as the task becomes more difficult, but nevertheless some can still withhold under certain conditions. Limitations, future directions, and implications for clinical practice are discussed.

Keywords: Child; Malingering; Cognitive Deficits.

Acknowledgements

I would like to thank my senior supervisor, Dr. Connolly, for her support, encouragement, and dedication throughout my graduate studies. I would also like to thank my committee for their advice and lending their expertise in conducting this research.

This dissertation would not have been possible without the support of my family, friends, and professional colleagues who had confidence in me and helped lift me up during the challenges one faces when completing a dissertation. I am forever grateful to all of you.

Table of Contents

List of Tables

Chapter 1.
Children's Ability to Malinger Cognitive Deficits

Holmes (1872) once said of children, "Pretty much all the honest truth-telling there is in the world is done by them." Parents, siblings, and teachers would call this naïve and argue that children have the ability to deceive. Although there is much research to support this (e.g., Evans & Lee, 2013; Talwar & Lee, 2002b; Talwar & Lee, 2008; Talwar, Murphy, & Lee, 2007), there is little research on whether children can deceive a psychologist during psychological testing that they have a head injury (Kirkwood, 2012). In the past, psychologists relied on clinical wisdom that children rarely fake head injuries on psychological tests (Salekin, Kubak, & Lee, 2008) and when children do fake, that it is easily detected (Kirkwood & Kirk, 2010). In the past decade, researchers and clinicians have begun to suspect that some parents may coach their children to fake head injuries, such as following a motor vehicle accident, to obtain financial compensation (see Sherman & Brooks, 2012 for a review). As a result, psychologists have become increasingly concerned about the possibility of pediatric malingering in assessing injury claimants (Kirkwood, 2015a; Sharland & Gfeller, 2007; Slick & Sherman, 2013).

The present set of studies seeks to address the following questions: (1) Can children between the ages of 6 and 12 deliberately withhold the correct answer on measures of memory and processing speed, and if so, at what age are they able to do so? (2) Are differences in children's ability to withhold related to their individual differences in inhibitory control? (3) How sensitive are the Test of Memory Malingering (TOMM) and Medical Symptom Validity Test (MSVT) at detecting children who are withholding the correct response?

1.1. Malingering & Detection

Definition. There are three ways that a person could present with physical or psychological symptoms without apparent causes: malingering, factitious disorder, or somatoform disorder. According to the DSM-5 (American Psychiatric Association, 2013), malingering is the intentional faking or exaggeration of symptoms to obtain external incentives, such as obtaining financial enrichment, avoiding work, or obtaining drugs. Slick, Tan, Sherman, and Strauss (2010) proposed four conceptual aspects of malingering for clinicians to consider when diagnosing malingering: There must be evidence that the individual (1) consciously chose to (2) expend less than optimal effort (3) towards achieving a short term goal of deception with the ultimate objective to (4) obtain a long term substantial personal gain, or avoidance of a personal duty or obligation.

As malingering has emerged in both research and clinical practice as an important topic, authors have used various terms to describe the effort exerted by examinees. Some have used the term “suboptimal effort”, however, Slick and Sherman (2012) noted that this may be a misnomer because level of effort may be unrelated to malingering. An examinee may exert little effort to feign a cognitive deficit using a simple strategy, or considerable effort if they are using a complicated strategy. For example, a simple strategy may be responding randomly, whereas an effortful strategy may require the examinee to exert effort above the demands of taking a test, such as keeping a strategy in mind, assessing individual item difficulty, deciding when to make errors, inhibiting the correct response, generating an alternative response, keeping track of performance over time to achieve a target error rate, monitoring the examiner’s reactions to their responses to assess the believability of their performance, and they may also be concurrently feigning symptoms such as fatigue, distress, or other physical symptoms (Slick & Sherman, 2012). It may be more relevant to consider whether the examinee is being compliant with test instructions to “try their best.” Thus, the results of an effort test actually indicate whether the individual is performing consistently with the instructions given. For this reason, throughout this dissertation, I will use the term “performance validity testing”, or “PVT”, although the term is generally synonymous with other authors who use the term “suboptimal effort.”

Similar to malingering, factitious disorder is the intentional production or feigning of physical or psychological signs/symptoms to assume the sick role. Unlike malingering, however, there are no external incentives. Rather, factitious disorder describes an individual whose primary motivation to dissimulate an injury is to satisfy a non-material psychological need, such as getting attention for having suffered an injury. Although there is some crossover between malingering and factitious disorder, especially when motivations are unclear, the salient differentiation is the source of the motivation. If a child feigns an injury in response to encouragement and instruction from a parent, or to obtain rewards/avoid consequences, this would fall within the realm of malingering. Conversely, if the individual is feigning an injury for his or her own internal incentive (i.e., psychologically motivated for attention), then it falls in the realm of factitious disorder. As it can be difficult to verify, there are no precise measures of the prevalence of factitious disorder among children. In one study of children and adolescents referred for psychiatric consultations in an Ontario hospital, the prevalence was estimated at 0.7%; however, the authors caution that this is likely an underestimate (Ehrlich, Pfeiffer, Salbach, Lenz, & Lehmkuhl, 2008).

Despite the lack of epidemiological data, there is clear evidence that factitious disorder does occur among children. Libow (2000) conducted a historical review of cases where it was thought that children feigned physical symptoms to assume the sick role and there was no parent involvement. Libow found 42 cases that ranged in severity from feigning a fever by warming a thermometer to actively injecting themselves to manipulate insulin levels. The cases included individuals ranging in age from 8-18 years old, and the majority (71%) were female. The average duration before the fabrication was uncovered was 16 months, which highlights the difficulty in detecting factitious disorder and provides an explanation for the likely understimate of its prevalence.

Unlike malingering and factitious disorder, an individual may present with invalid symptoms without an apparent motivation to deceive. In such cases, it is possible that the individual has a somatoform disorder in which they report suffering from a physical illness yet there is no physical medical explanation for their symptoms. For these individuals their symptoms are caused or exacerbated by involuntary psychological factors such as stress.

Although malingering, factitious disorder, and somatoform disorder are fairly rare, one must consider the possibility of their presence when an injury appears invalid. This

dissertation focuses exclusively on malingering; factitious and somatoform disorders will not be discussed further.

Prevalence of malingering. Estimating the prevalence rate of malingering is difficult as the nature of malingering is that the individual is trying to conceal their intent. Among U.S. and Canadian adults who present for personal injury, disability, criminal, and medical litigation cases, neuropsychologists estimate that around 30-40 % of examinees are suspected of exaggerating or faking impairments (Sharland & Gfeller, 2007). Unfortunately, a limitation to this survey is that it only included a small proportion of survey respondents from Canada and there is little information on whether there are differences in the prevalence rates of suspected malingerers between the U.S. and Canada (Mittenberg, Patton, Canyock, & Condit, 2002).

Larrabee (2003) calculated the base rates of malingering from 11 studies that used objective measures of malingering with adults (n = 1363). He found the base rate of malingering mild head injuries in the context of potential secondary gain to be approximately 40%. As a result of the high rate of malingering in some contexts, the National Academy of Neuropsychology has recommended that response validity be assessed in order to maximize the confidence in neuropsychological test results.

As a result of the growing awareness of the prevalence of dissimulation in neuropsychological assessments, neuropsychologists have become increasingly aware of the need to assess performance validity, particularly when secondary gain is possible, such as in the context of personal injury and/or criminal litigation (Boone, 2007). In brief, neuropsychologists use specific tests or scales embedded into other tests that detect performance validity. These tests or scales appear to measure an ability, such as memory, but the tests are generally insensitive to legitimate impairment. Thus, performance below a specific cut-off alerts the assessor that the examinee may not be following the instructions to try their best, or in other words, their performance may be invalid. These tests are limited in that they are not perfectly sensitive or specific, and they do not provide any explanation for why the individual may have scored below the cut-off, such as genuine severe impairment, fatigue, or language deficits.

Sharland and Gfeller (2007) reported a decade ago that 57% of the neuropsychologists in the U.S. and Canada who responded to their survey indicated that they frequently include performance validity tests (PVTs) in their forensic assessments. Since that time, there has been a paradigm shift in neuropsychologists' approach to

forensic assessments. In a more recent survey, nearly all neuropsychologists surveyed reported that it is *essential* to include PVTs in forensic assessments and that it is desirable in clinical assessments (Martin, Schroeder, & Odland, 2015). Essentially, it has become standard practice for neuropsychologists to include tests to detect performance validity in their assessments.

Data on the prevalence rate of malingering among children is limited. It has been suggested that the lack of epidemiological data may have resulted from the historical view that children are honest and do not have the capacity to deceive (Salekin et al., 2008). Studies attempting to estimate the prevalence rate of malingering use known groups, such as individuals undergoing assessments for personal injury litigation. The performance of children in known groups is compared to groups of individuals who are being assessed for purposes where there is little or no potential for secondary gain, such as diagnostic assessments for treatment, or control groups where individuals are simply asked to try their best.

In a clinical study in which children were assessed for social security disability benefits, a situation with potential for parents' secondary gain, children were administered two PVTs, the TOMM and the MSVT. Twenty-eight percent of the children failed the TOMM and 37% failed the MSVT (Chafetz, Abrahams, & Kohlmaier, 2007). In another study, Kirkwood and Kirk (2010) examined a group of mostly non-litigating children with mild traumatic brain injuries who were referred for neuropsychological evaluations and found that 17% failed the MSVT. If we consider that almost all children tend to pass PVTs when they are trying their best, then it appears that a sizable number of examinees were likely not performing optimally when secondary gain was possible.

An important consideration from these studies is that secondary gain is not limited to the context of personal injury litigation. The Kirkwood and Kirk (2010) study included mostly non-litigating children and the Chafetz et al. (2007) study included children being assessed for disability benefits, which would presumably go to the child's parents. There are also case studies, such as the one described by Conti (2004), where children have malingered ADHD to obtain stimulant medication to sell to other children. In addition, it is not inconceivable that children being assessed for educational purposes may perform suboptimally on achievement testing to obtain academic accommodations, such as more time to write tests or the use of technology aids to make exam writing easier.

In summary, research has shown that as much as 40% of adults making claims for personal injury, disability, or medical claims may be exaggerating or fabricating their symptoms to obtain compensation. As a result, neuropsychologists routinely include tests to assess performance validity in their test batteries, especially when there is a potential for secondary gain, such as financial compensation following an injury. While there has been less research on the prevalence rate of pediatric malingering than adult malingering, there is evidence that children may feign their symptoms in particular contexts as well. When cognitive injury claims are adjudicated within the legal arena, the courts have established parameters for the compensability of cognitive and psychiatric injuries. Most often, these types of injuries fall within tort law, which I will discuss next.

1.2. Tort Law

Tort law is an area of law that is intended to correct injustices caused by the wrongful conduct of the defendant (Linden, 2015). It is primarily intended to serve the function of compensation of losses to the plaintiff to return them to the condition they would be in had the tort not occurred. Although compensation is the primary function of tort law, it also serves, in limited circumstances, the function of punishment, deterrence, accountability, corrective justice, appeasement, vengeance, and education (Osborne, 2015). Although there are various specific torts, they all share similar principles and are defined by the nature of the defendant's conduct and the plaintiff's loss resulting from that conduct (Fleming, 1998). Most often, claims regarding cognitive injuries are adjudicated in this area of law.

For a plaintiff to make a successful tort claim, causation must be established by satisfying the court that "but for" the defendant's act, the plaintiff would not have sustained their injury. This was most recently affirmed in Clements v. Clements, 2012 SCC 32.

Tort Law is generally divided into two categories: Intentional torts and negligence. Intentional torts are defined by the intentional conduct of the defendant, such that the defendant *intended* to cause harm to the plaintiff (e.g. an assault). In negligence, intentionality is not required and the defendant may be liable for their conduct if a "reasonable person" ought to have foreseen the consequence of their

behaviour (e.g. driving recklessly in a school zone). The latter type of tort is more common, although both may result in compensable psychiatric injuries (Koch, Douglas, Nicholls & O'Neill, 2005).

Although the courts now recognize psychiatric injuries as a distinct form of injury, this has not always been the case. Osborne (2015) provides four potential reasons why the Courts have historically approached the compensation of psychiatric and cognitive injuries cautiously and conservatively. These policy considerations include: (1) fear of opening the floodgates, (2) difficulty linking the defendant's conduct to the cognitive and/or psychiatric injury, (3) lack of empathy, and (4) the perception that cognitive and psychiatric injuries can be easily fabricated (Osborne, 2015). First, the fear of opening the floodgates to an indeterminate number of plaintiffs is one of the most frequently cited reasons for limiting liability. The concern is that a large number of plaintiffs would overwhelm the judicial system and place an undue burden on the defendant that is grossly disproportionate to the degree of their wrongful conduct (Osborne, 2015). Second, the courts may be reluctant to award damages because it can be difficult to determine whether the cognitive and/or psychiatric injury can be attributed to the defendant's conduct or some other life experience. Unlike a visible physical injury, such as a broken arm, the causal nexus between the defendant's act and the plaintiff's injury may not be readily apparent and the resulting injury may only become evident after a period of time has passed. Indeed, cognitive and psychiatric injuries are only considered sufficiently severe when a period of time has passed and the symptoms have not remitted. This intervening time introduces the possibility that another event in the plaintiff's life caused or exacerbated the injury (Osborne, 2015). Third, the courts may be reluctant to recognize psychiatric injuries because of the stigma associated with mental illness and the associated lack of empathy towards those with psychiatric injuries. While everyone can recognize and appreciate the pain of visible injuries, psychiatric injuries are less familiar and may elicit less empathy. Lastly, it is sometimes claimed that cognitive and psychiatric injuries are more easily fabricated than physical injuries and they should therefore be treated skeptically (Osborne, 2015). Unlike evidence available from an x-ray or visible wound, cognitive and psychiatric injuries are not directly observable and may be based on self-report. Although they still remain difficult to verify with absolute certainty, psychological tests have been devised to aid in assessing the veracity of cognitive or psychiatric injury claims. These tests have proved promising to

detect dissimulation, however, evidence of their applicability to certain populations, such as children, is lacking. Collectively, these factors have historically contribute to the high threshold required for, and limited conditions under which, damages for cognitive or psychiatric injuries may be obtained.

Damages. One of the key features of determining liability is whether a plaintiff can demonstrate that they have suffered an injury. The resulting damage from a defendant's behaviour is key to triggering a tort claim and is necessary to assigning liability to the defendant. The legal system, however, cannot compensate everyone for every loss resulting from intentional or negligent conduct, as this would place an unfair burden on the defendant and their insurers. As a result, tort law must define what is compensable and what is not, and apply control devices to create acceptable limits on liability.

Drawing this line is particularly difficult in cases of cognitive or psychiatric injury. The case of *Hinz v. Berry* (1970) is illustrative. In this case, a family was parked on the side of a country road to have a picnic when the mother and one of her children crossed the road to pick flowers. While they were picking flowers, a car driven negligently by the defendant crashed into the family car, killing the father and injuring the other children. The family's claim for financial support and the children's claim for their physical injuries were settled without the need for a trial. The mother also sought compensation for nervous shock from witnessing the accident and other forms of mental distress including grief, sorrow, depression, anxiety about her children's welfare, and financial stress. The Court only granted compensation for her depression because it is a recognized psychiatric illness and was directly attributable to the shock of the accident and not a subsequent consequence of her losses.

Determining whether a plaintiff's psychiatric injury is attributable to a defendant's conduct may be exceedingly difficult due to pre-existing or concurrent stressors, or intervening events between when the accident occurred and when the psychiatric injury became evident. In the case of *Birkich v. Cantafio* (2016) the plaintiff was a young woman who had experienced many psychosocial stressors growing up, which put her at risk for developing mental health problems. At age 17, she was struck by a truck and suffered physical injuries as well as a concussion. Following the accident, the plaintiff suffered several unrelated stressors, including familial discord, an unplanned pregnancy, and a relationship breakup. She concurrently developed depression and claimed that it

was a result of her injury. The defendant did not challenge the plaintiff's claim of having suffered a concussion, but argued that her mental status was a result of her pre-existing vulnerabilities and the stressors that arose following the accident. Moreover, the defendant argued that the victim's cognitive complaints could be explained by her mental health issues rather than by her concussion. In essence, the defendant argued that "the plaintiff would ultimately have suffered from the problems, even if the motor vehicle accident had not occurred." Justice Betton disagreed and concluded that despite the plaintiff's vulnerabilities, "The incident was in fact a significant traumatic event that did result in a mTBI" (mild traumatic brain injury) which initiated the plaintiff's symptoms. He stated that the plaintiff's cognitive and psychological symptoms were "indivisible" and while they may be interrelated, they were nevertheless triggered by the defendant's conduct.

Although the general trend is to only compensate for psychiatric illnesses that meet diagnostic criteria (see *Odhavji Estate v. Woodhouse*, 2003), there are instances where the courts have allowed claims for psychological distress that did not meet the threshold for a diagnosis. In *Anderson v. Wilson* (1999), the Ontario Court of Appeal reversed a decision to exclude plaintiffs in a class action suit that did not meet diagnostic criteria for having a psychiatric illness. In his reasons, Justice Carthy cited two Ontario Provincial Court cases that awarded damages for mental distress that did not meet criteria for a diagnosis. In *Mason v. Westside Cemeteries Ltd.* (1996), damages were awarded for emotional upset to a son when the funeral home lost the cremated remains of his parents and in *Vanek v. Great Atlantic & Pacific Co. of Canada* (1999), Judge Cosgrove awarded damages for subclinical anxiety and stress to a father whose daughter drank contaminated juice. Similarly, in the BC Supreme Court case of *McDermott v. Ramadanovic Estate* (1988), Justice Southin awarded damages to a young girl who suffered the "emotional scar" of witnessing her parents' death in a motor vehicle accident, even though her emotional distress did not amount to a recognized psychiatric illness. Justice Southin drew a parallel between physical and emotional scars and concluded that emotional scars are no different and should therefore be compensable.

While some judges have used the criteria that a psychiatric injury must be "serious and prolonged" and not "minor or transient", others have applied control mechanisms to maintain fairness between plaintiffs and defendants. In *Kotai v. Queen of*

the North (2009), Justice Joyce noted that it is problematic to formulate an objective measurement of “serious” and unclear on how to measure emotional impact, as it is inherently subjective. Furthermore, he asked whether the emotional distress would have to impact the plaintiff’s emotional state or create some other functional impairment. Justice Joyce argued that the *recognized psychiatric illness* standard introduces objectivity by allowing expert witnesses to provide evidence and a clear benchmark for the courts.

In other cases, Judges have considered the severity of the psychiatric injury, rather than the binary presence or absence of a diagnosis, or the functional impairment resulting from the psychiatric injury. In Y*oshikawa v. Yu* (BCCA 1996), Justices Lambert, Cumming, and Rowles outlined various criteria that must be met for a psychiatric injury to be compensable. These include, but are not limited to: the pain, discomfort, or weakness must be genuine; the psychological problems must have their cause in the defendant’s wrongful act and not be rooted in desires for sympathy or compensation or be such that the plaintiff could be expected to overcome them through his or her inherent resources; the psychological mechanism is beyond the plaintiff’s power to control and was set in motion by the defendant’s wrongful act; and that the evidence of psychological problems must be of a “convincing nature” but the plaintiff’s own evidence, if consistent with the surrounding circumstances, may suffice for the purpose. It is noteworthy that the presence of a formal diagnosis is not required in these criteria, nor does an expert have to confirm the presence of a diagnosis. In *Saadati v. Moorhead* (2015), however, the absence of a diagnosis was used as the basis for an appeal. In this case, the plaintiff was involved in a motor vehicle accident. He appeared to be concussed following the accident and reported symptoms associated with a concussion to his physician a few days later. During the trial, the plaintiff’s friends and family provided evidence that following the accident, the plaintiff was irritable, cognitively slow, and less charming than before the accident. The trial judge awarded the plaintiff $100,000 in non-pecuniary damages based solely on the plaintiff’s family and friends reporting changes in him following the accident. On appeal, however, Justice Frankel overturned the decision as the plaintiff had not proven that he suffered a recognizable psychiatric illness as a result of the injury.

In cases of cognitive injury, courts have accepted that brain injuries may be difficult to identify and the severity of an injury must be judged on the evidence as a

whole (Christoffersen, 2013). This may include a witness' observations of the individual's behaviour following the accident (e.g., loss of consciousness, confusion, disorientation), reported changes in work and cognitive functioning (e.g., decline in memory, concentration), self-reported symptoms (e.g., dizziness, headaches), medical reports (e.g., MRI results), and performance on neuropsychological testing. If a court is convinced that a cognitive injury is present and the individual's recovery has reached a plateau, the court must then determine the appropriate remedy.

Remedies. The law allows for two types of damages to be awarded: pecuniary and non-pecuniary. Pecuniary damages are intended to compensate the individual for the financial loss resulting from the injury and can be subcategorized as special damages (pecuniary costs incurred pre-trial such as medical expenses) and general damages (pecuniary damages that will be incurred post trial such as future loss of earnings). The guiding principle for awarding damages in Canada is to provide financial compensation to return the victim to their pre-tort condition (Linden, 2015). Non-pecuniary damages are intended to compensate for non-tangible losses, such as pain and suffering, and loss of enjoyment in life. The purpose of awarding non-pecuniary damages is to provide the injured person with money to purchase goods and services that will give them solace for their loss (see *Andrews v. Grand and Toy Albert Ltd.,* 1978). Determining the extent of both types of losses may be dependent on the court's judgment of the plaintiff's credibility. In cases where this is an issue, performance on neuropsychological testing may be instrumental in determining whether the plaintiff is making a legitimate and truthful claim.

Once a compensable injury has been established, defendants are generally liable for the reasonably foreseeable consequences of their conduct. They are also fully responsible for restoring the plaintiff to their pre-tort condition, to the extent that money can do this. The thin-skull rule (sometimes called an "eggshell personality" in psychiatric injury cases), dictates that the defendant remains liable for the injury they caused even if the victim suffered a more severe injury than would have been expected for an average reasonably resilient person. In other words, the defendant must take their victim as they find them. This has been limited, however, in cases where the psychiatric damage suffered by the plaintiff was not reasonably foreseeable. In *Mustapha v. Culligan of Canada* (2008), the plaintiff found a dead fly and parts of a dead fly in his bottled water and claimed that as a result, he could no longer drink water, shower, or have sex. The

plaintiff also developed major depressive disorder which impaired his daily functioning. The Supreme Court of Canada affirmed that the defendant's reaction was beyond what would be expected of a "reasonable person" and the defendant could not have foreseen that the defendant would react in this way. As a result, the defendant was not held liable for the plaintiff's psychiatric injury.

Another limitation to a defendant's liability is that they are not responsible for returning their victims to a better condition than they found them. If the plaintiff had a pre-existing condition that was exacerbated by the defendant's conduct, termed a "crumbling skull," the defendant remains liable but there may be a reduction in the damages awarded to the plaintiff (Blackwater, 2005; Koch, Douglas, Nicholls, & O'Neil, 2006). This circumstance may apply in cases of cognitive injury where the defendant's cognitive status prior to the injury had an impact on their functioning. This raises a challenge for assessors to estimate pre-morbid functioning and then provide an opinion on the degree of cognitive impairment resulting from the injury.

1.3. Detecting Deception

As discussed above, the courts have approached psychiatric injuries with caution because these types of injuries cannot be visibly observed like an x-ray (although the behaviour resulting from a psychiatric injury may be observable) and there is a belief that psychiatric injuries can be easily fabricated. Triers of fact may have to evaluate whether a plaintiff's claim of having suffered a psychiatric injury is credible by attempting to evaluate whether they are lying. This has proved to be extraordinarily challenging with most empirical studies showing that liars are difficult to detect. In this section, I review the accuracy rates of which adults' and children's lies are detected and under what conditions, as well as some of the systematic methods that have been used to detect lies.

Detecting adults' lies. Historically, it was thought that deception could be detected using observation and intuitive judgment alone. Studies examining this assumption refer to people who make deceptive or truthful statements as *senders* and people who judge the statements as *receivers*. The statements made by the senders are called *messages*. In most studies, receivers judge whether the message is truthful or

deceptive, and thus chance performance is 50%. In one of the earliest reviews of the accuracy of deceptive judgments, Kraut (1980) found a mean accuracy rate of 57% across ten studies, indicating that individuals are only slightly better than chance at detecting lies. Twenty years later, Vrij (2000) conducted a similar review with a larger and updated sample of studies. He replicated the findings of the previous review, finding that individuals achieved accuracy rates of 56.6%. The consistency between these findings illustrates that people, when left to detect deception without aids, make poor lie detectors.

C. F. Bond and DePaulo (2006) conducted a large-scale meta-analysis including 206 studies and 24,483 receiver judgments. In this analysis, the authors sought to examine lie detection accuracy more closely by examining additional variables such as the medium by which the message is sent, the liar's motivation, whether the sender was given a chance to prepare their lies, whether the receiver was provided baseline exposure to the sender, whether there was interaction between the sender and receiver, and the receiver's level of "expertise" to detect lies. Overall results showed a mean accuracy rate of 54%. Similar to previous reviews (e.g., Vrij, 2000), the analysis also found a bias to judge messages as truthful. This resulted in a slightly higher accuracy in truthful judgments (61.3% accuracy) than deceptive judgments (47.6% accuracy) that could be attributed to bias rather than accurate detection. The presentation medium made a significant difference in accuracy judgments; video only presentations yielded lower accuracy rates than audiovisual, audio only, and transcript presentations. Accuracy rates from transcript, audiovisual, and audio only presentations were not significantly different than one another. The degree of sender motivation to deceive impacted the accuracy of receiver judgments. Lies were easier to discriminate from truths when told by a motivated sender compared to an unmotivated sender. C. F. Bond and DePaulo (2006) suggest that the efforts of motivated senders undermine their ability to successfully deceive. When senders were allowed time to prepare their messages, receivers were less accurate at detecting deception compared to spontaneous messages. Allowing receivers baseline exposure to senders also improved the accuracy of judgments. Receivers with baseline exposure to the senders obtained a mean accuracy rate of 55.9%, whereas those without baseline exposure obtained a mean accuracy rate of 52.3%. It was hypothesized that senders would have more difficulty deceiving the receiver with the increased demands of the social interaction, however,

receivers were no better at detecting deception in the context of an interaction than merely observing the sender. When lies were in the form of a narrative, receivers were unreliable and inaccurate at differentiation between lie-tellers and truth-tellers. In addition, receivers were no better at detecting liars when the lie was a short response ("Did you cheat on the test?") or longer simulated narrative (i.e., a story about a false event). In sum, while there were variations in receivers' accuracy rates under some circumstances, performance was never impressive.

Receivers who are required to detect deception as part of their occupation, such as police officers, customs officials, and judges were no better at detecting deception than laypersons, which were often college students. These "experts" tended to be more skeptical than laypersons and more likely to judge senders as liars than laypersons. Despite their poor performance, however, "experts" are frequently more confident in their abilities than laypersons and overstate their accuracy (Frank & Ekman, 1997; Leach, Talwar, Lee, Bala, & Lindsay, 2004).

It has been suggested that there are "wizards" of lie detection who achieve very high rates of deception detection (Ekman & O'Sullivan, 1991; Ekman, O'Sullivan, & Frank, 1999; O'Sullivan, 2007). G. D. Bond (2008) assessed the accuracy of two individuals with apparent expertise and found that they achieved 80-90% accuracy rates by relying on non-verbal cues. Similarly, Ekman and O'Sullivan (1991) found that a sample of 34 Secret Service agents were over 70% accurate in evaluating truthful and deceptive statements made by undergraduate students. In sum, although there may be some experts who are particularly skilled at detecting lies, the majority of the research suggests that individuals and groups who achieve better than chance accuracy rates have done so by chance alone.

Some have argued that lie detecting is an ability that varies between individuals (O'Sullivan, Frank, Hurley, & Tiwana, 2009). C. F. Bond and DePaulo (2008) examined large distributions of individuals' performance at detecting deception and found that individuals who achieved accuracy rates that were above chance levels were still within the range of expected random variability. In other words, although some individuals achieved higher than chance accuracy rates, the number of individuals at the end of the spectrum was within the range of what would be expected by random variation alone. Rather than focusing on the abilities of the receiver, C. F. Bond and DePaulo (2008) suggest that senders vary in their ability to deceive. They note that some individuals

appear quite adept at telling convincing lies whereas others are more easily detected. They found that individual differences in sender ability to evade detection accounts for more variability than receiver's detection ability.

Detecting children's lies. Compared to the literature on detecting adults' lies, there are relatively few studies that have examined adults' ability to detect children's lies. The majority of these studies show that adults have difficulty detecting children's lies (Chahal & Cassidy, 1995; Edelstein, Luten, Ekman, & Goodman, 2006; Leach et al., 2004; Newton, Reddy, & Bull, 2000; Oldershaw & Bagby, 1997; Westcott, Davies, & Clifford, 1991; Wilson, Smith, & Ross, 2003), although there are differences based on the child's age (Feldman, Jenkins, & Popoola, 1979; Westcott et al., 1991), whether the child was engaged in a moral discussion before questioning (Leach et al., 2004), and the type of lie (Talwar, Lee, Bala, & Lindsey, 2004).

The research on whether adults are better able to detect the lies of younger versus older children is unclear. In one study (Westcott et al., 1991), adults evaluated the truthfulness of children 7-11 years old talking about a visit to a history museum. Half of the children had visited the museum and half had watched a video about the museum and were instructed to answer the interview questions as if they had visited the museum. Adults were most accurate in detecting the truthfulness of 7-8 year old boys (78.5% accurate) and performed at chance for detecting the truthfulness of young girls and both older boys and girls. In a similar study (Talwar, Lee, Bala, & Lindsay, 2006), children age 4-7 were either asked to tell a story about a personally experienced event or they were coached by their parents to tell a fictional story about an event that never happened. Adults were unable to distinguish truthful from deceptive stories and boys were more often judged as liars than girls. The results of these two studies suggest that adults are poor detectors of children's lies and may have a gender bias to judge boys as more likely to lie than girls.

Leach et al. (2004) examined whether engaging children (age 3-11) in a moral discussion or having them promise to tell the truth influenced their presentation and subsequent detectability as liars. In the moral reasoning condition, children were engaged in a discussion about the importance of telling the truth. Children in the promise condition were asked to promise to tell the truth. It was hypothesized that increasing children's awareness of the importance of telling the truth would increase their negative affect associated with lying and make their deception easier to detect. For both the moral

reasoning and promise conditions, adults were better able to discriminate truthful from deceptive statements than would have been predicted by chance, although lie detection never exceeded 70%. There was a trend that adults were slightly better at judging younger than older children. When adults evaluated children's truthful or deceptive statements without the moral discussion or promise, they performed at chance levels.

Similar to findings of a truth bias in judgments of adults, there was also a tendency to evaluate children as truthful. Talwar et al. (2006) found that adults judged the majority of children age 4-7 as making truthful statements, yielding a higher accuracy rate for truthful statements than deceptive statements. Similarly, Evans, Bender, & Lee (2016) found that adults have a truth bias in evaluating children age 8-16. Interestingly, parents evaluating their own children's statements had an even stronger truth bias than parents who were evaluating other children's statements. Regardless of the relationship, however, adults performed at chance when judging the veracity of statements. Thus, a close relationship may hinder an adult's ability to detect a child's lies. The finding that parents have difficulty detecting their own children's lies has also been replicated among parents of younger children. Talwar, Renaud, & Conway (2015) asked parents to detect when their own child (age 3-11) was lying. Similar to other studies involving strangers, parents were only slightly better than chance at correctly identifying when their child had lied and had greater difficulty detecting older children's lies than younger children's lies.

There is evidence that children age 3-11 are better able to lie convincingly when motivated to conceal a transgression such as not peeking at a toy (e.g., Talwar et al., 2004), however, these lies tend to be simple and only require the child to give brief responses to questions (e.g., "Did you peek at the toy?" – "No"). When young children are required to respond to follow-up questions and maintain their lie in subsequent questioning, they tend to have difficulty maintaining their lie in a way that is consistent with their earlier simple lie. For example, in the study by Talwar and Lee (2002a), children age 3-7 were asked about the type of toy they initially denied peeking at. The majority of the children responded correctly (i.e., they identified the toy they had just reported not peeking at) thereby implicating themselves as liars. This inability to provide consistent responses to an initial lie is called *semantic leakage*. Some of the older children in the sample were more sophisticated in their responses to subsequent questioning and exhibited greater control over their semantic leakage.

The research shows that although there are individuals who believe they can detect liars, there are very few with accuracy rates above chance. As a result, individuals have turned to technology, such as physiological measures and neuropsychological testing to assist with detecting lies. A discussion of physiological measures is beyond the scope of this dissertation, but interested readers can refer to Vrij & Fisher, 2016 for a review of the research and *R. v. Beland* (SCC 1987) for the decision that polygraph results cannot be introduced because it does not meet the standards of the rules of evidence in Canada.

Structured assessments of children's statements. Statement Validity Assessment (SVA) is a forensic technique to assess the veracity of children's statements. It is mostly used to assess statements about sexual abuse, but it has also been used for other purposes (e.g., Tye, Amato, Honts, Devitt, & Peters, 1999). At the core of SVA is Criteria-Based Content Analysis (CBCA) which has been the focus of research. The principle underlying the technique is based on the Undeutsh hypothesis, which states that recollection of a self-experienced event will be different in content and quality from the recollection of a fabricated or imagined event (Undeutsch, 1989). The procedure involves the systematic evaluation of 19 criteria that are considered characteristic of truthful accounts. There are three categories of criteria and an account is considered likely to be true if a high number of criteria are present (Blandón-Gitlin, Pezdek, Lindsay, & Hagen, 2009), however, there is an allowance for the evaluator to take into consideration other factors that could have affected the outcome, such as the child's cognitive abilities. Various studies have found support for the Undeutsch principle, but there are inconsistent findings on the optimal score to differentiate between truthful from fabricated accounts (Vrij, 2006). The method has been criticized for being sensitive to coaching (Vrij, Akehurst, Soukara, & Bull, 2002) and influenced by the sender's familiarity with the event (Blandón-Gitlin, Pezdek, Rogers, & Brodie, 2005). In addition, the empirical support from field studies is limited as the actual veracity of children's statements is unknown (Vrij, 2005). Despite these limitations, Vrij (2005) found that accuracy rates from lab-based studies of CBCA were 73% for truthful statements and 72% accurate for lies, and overall accuracy rates from independent studies ranged from 65-90% accurate.

Structured psychological tests. Structured psychological tests to detect deception have garnered much attention over the past three decades and are now commonplace in neuropsychological and forensic assessments. These tests are used in evaluations to detect whether an examinee may be feigning symptoms within a psychological evaluation and cannot be used to detect everyday lies. Historically, psychologists did not suspect that examinees would not follow the test instructions to try their best during formal assessments. In an often cited study from 1988, Faust, Hart, and Guilmette asked three adolescents (age 15-17) to perform less well than their capabilities on the Wechsler Intelligence Scale for Children-Revised (WISC-R), Halstead-Reitan Neuropsychological Test Battery, Aphasia Screen, Grip Strength, Reitan-Klove Tactile Form Recognition, Suppressions on Simultaneous Stimulation, Tactile Finger Recognition, Finger Tip Writing, and Wechsler Memory (Russel Adaptation). Symptoms validity tests were not included in the battery. Adolescents were instructed to *not* try their best, but to be convincing enough that they are not caught as faking. Compared to an actual head injury test profile (one of the control conditions), the simulators showed an unusual test profile with large discrepancies between abilities (e.g., a 20 point Verbal Intelligence Quotient – Performance Intelligence Quotient split on the Wechsler Adult Intelligence Scale – Revised). Three hundred neuropsychologists were asked to review the test profiles and choose between three choices to explain the results: cortical dysfunction, malingering, or functional factors. Ninety-three percent of the respondents identified the profile as abnormal, yet nearly all of them suspected cortical dysfunction and none suspected malingering. Despite the poor performance at detecting deception, three quarters of the respondents indicated moderate or greater confidence in their judgment (Faust et al., 1988; Faust, Hart, Guilmette, & Arkes, 1988). Although this study can be criticized for not providing neuropsychologists with sufficient supplemental information to evaluate malingering, such as consistency of information (Bigler, 1990) or scores on PVTs, it does show that neuropsychologists in the late 1980's did not suspect malingering based on these atypical test profiles, and that relying on clinical judgment alone is fallible. With recent increases in the use of PVTs and an increased awareness of the prevalence of malingering, it is not known whether these results would be replicated today.

To address the issue of malingering with adults, psychologists have developed PVTs to assess whether an individual may be deliberately performing poorly (Sharland &

Gfeller, 2007). There is substantial experimental and clinical support for the use of these tests (Bush et al., 2005) and they are commonly included in test batteries for individuals within populations where secondary gain is possible, for instance among personal injury litigants (Slick, Tan, Strauss, & Hultsch, 2004).

The principle behind using PVTs is that failure on at least one PVT is associated with a decrease in scores across multiple cognitive domains and reflects the examinee's general approach to testing (Kirkwood, Yeates, Randolph, & Kirk, 2012). Scores on PVTs are not intended to be interpreted as directly corresponding to whether an individual is malingering or not. Clinicians use a multi-test and multi-method approach to assess the examinee's approach to testing. PVTs provide useful information, but it must be considered within the broader context that includes interview data, mental status, behavioral observations, collateral information, the individual's history, neuropsychological test performance, and any other available information (Walker, 2011).

There are two types of tests that can be used to assess performance validity: indices from conventional tests and tests specifically designed to detect performance validity (Strauss, Sherman, & Spreen, 2006). Both types of tests are used to assess whether performance is consistent with the expected performance by individuals with particular neurological problems. For conventional tests, examiners use empirically validated cut-off scores or look for atypical patterns of scores to look for possible suboptimal performance. If an individual scores far below what would be expected for a clinical population, or performs differently on two subtests that measure the same ability, it may be indicative of malingering (Strauss et al., 2006). The second type of test is specifically designed and intended to measure performance validity and is sensitive to performance validity but not to genuine impairment. These tests rely on probabilities that an examinee who is not performing optimally will score below chance, or lower than would be expected compared to individuals with genuine injuries. For example, if an examinee is given a forced choice test with two response options per item, one would not expect them to perform lower than approximately 50%, which would be chance performance. Similarly, if an individual scores below average on a test that is not sensitive to injury (i.e., both healthy and injured test takers perform equally well when following the instructions to try their best), it *may* be indicative of suboptimal performance. Importantly, a low or inconsistent score may not only indicate malingering

or symptoms exaggeration; there are a number of reasons why an individual may fail a PVT, such as low cognitive ability, insufficient language ability, attention/concentration problems, age, severe neurological disorder, and active psychiatric illness (Greiffenstein, Baker, & Gola, 1994).

One way to reduce the false positive rate (incorrectly identifying an individual as performing suboptimally) is by considering other test data to determine whether the severity of the deficit accounts for the failure on the PVT. Green and colleagues (Green, Flaro, Brockhaus, & Montijo, 2012) used a profile analysis strategy to reduce the rate of false positives in a sample of children referred for assessments with severe cognitive impairments. For children who failed the MSVT, they compared the easy subtest scores (IR, DR and CNS scales) to the more challenging subtest scores (PA and FR). If the individual obtained a score on the PA or FR scales that was at least 30-points lower than the scores on the IR, DR, or CNS scales, the profile was classified as reflecting genuine impairment and not a false positive. If the difference was less than 30-points, the profile was suggestive of suboptimal effort. Using this profile analysis strategy, the authors reduced the rate of false positives from 21.6% to 3.3% (96.7% specificity). Larochette and Harrison (2012) applied the same profile analysis strategy to a sample of adolescents with severe learning or reading disabilities using the Word Memory Test (a longer version of the MSVT) and were able to reduce the false positive rate from 10% to 1%.

It is generally recommended that examiners administer at least two measures to assess performance validity in their assessments to increase confidence in their decision-making (Boone, 2007; Boone & Lu, 2003; Larrabee, 2005). This recommendation was supported by Larrabee (2008) who administered multiple PVTs to individuals and found that multiple tests increased the likelihood of correctly identifying an examinee as malingering. Based on his data, Larrabee (2008) recommends using three PVTs as doing so achieved nearly perfect classification accuracy with his sample. The specific number of tests to use, however, depends on the base rate of malingering, the psychometric properties of the test, the type of tests, and whether the tests are redundant or provide incremental utility (Kirk, Hutaff-Lee, Connery, Baker, & Kirkwood, 2014). It is worth noting that failure on one PVT and not another may also reflect variable performance validity throughout the testing session. Thus, clinicians should also consider the placement of their PVTs within their test battery.

There is discussion in the literature on how clinicians should combine results from validity tests with other data. Chafetz and colleagues (2007) have suggested using a rating scale called the Symptom Validity Scale (SVS) which combines performance on quantitative scales, such as embedded measures of performance validity like the Reliable Digit Span, with qualitative aspects of test performance, such as not knowing autobiographical information or providing other atypical responses. In his studies, the SVS was validated against the MSVT and TOMM. The SVS, which includes various parameters, was found to be more predictive of performance validity than any single parameter, such as failure on a specific PVT.

In addition to using multiple measures of performance validity, neuropsychologists typically use multiple methods to assess malingering. In a survey of neuropsychologists, Sharland and Gfeller (2007) identified some of the most commonly used methods, which included: comparing the severity of the cognitive impairment to the severity of the injury, comparing medical records to self-report and observed behaviour, assessing whether the pattern of cognitive impairment is consistent with the condition, assessing for implausible self-reported symptoms, and comparing test scores across repeated examinations (Sharland & Gfeller, 2007). Thus, PVTs are not to be used in isolation and are not intended to be the only source of information in making a determination of malingering.

Of the neuropsychologists who responded to the 2007 survey, the TOMM was the most commonly used PVT (25% reported always using it). Consistent with their clinical practices, survey respondents indicated that they believed the TOMM was the test best able to correctly classify adequate from inadequate effort (Sharland & Gfeller, 2007). In a similar survey of forensic psychologists, the TOMM was rated as the most commonly used PVT and the second most commonly used test of symptom malingering after the Structured Interview of Reported Symptoms (SIRS; Archer, Buffington-Vollum, Stredny, & Handel, 2006).

The TOMM is a test that is specifically designed to assess performance validity. It appears to measure memory, although individuals with organic cognitive and memory deficits (such as those with traumatic brain injury, aphasia, cognitive impairment, and dementia) are able to complete the task to ceiling or near ceiling accuracy (Hill, Ryan, Kennedy, & Malamut, 2003; Tombaugh, 1997). It has been noted, however, that examinees with severe cognitive impairment and moderate to severe dementia are

occasionally misclassified as malingering (Rees, Tombaugh, Gansler, & Moczynski, 1998; Strauss et al., 2006). The test has two learning trials and a delayed retention trial. The maximum score is 50/50 and any score less than 45 on trial 2 should raise concern that the individual may be performing suboptimally in an attempt to malinger. Using the recommended cut-off, the specificity of the test exceeded 90% for a group of impaired adults, including those with genuine traumatic brain injuries (TBI) and various neurological conditions (Tombaugh, 1997). Donders (2005) found that performance on the TOMM is unrelated to the length of time the injured person was in a coma following their injury, which is related to the severity of the trauma. Moreover, the TOMM is a useful PVT because examinees tend to have difficulty identifying it as such a test (Rees et al., 1998). Another advantage of the TOMM is that the stimuli to remember are pictures, and thus the test is less sensitive to language difficulties. Although the TOMM has been less extensively researched with children as it has been with adults, studies indicate it can be used reliably with children as young as 5 years old (Blaskewitz, Merten, & Kathmann, 2008; Constantinou & McCaffrey, 2003; Donders, 2005; DeRight & Carone, 2015; Gast & Hart, 2010; Gunn, Batchelor, & Jones, 2010; Kirk, Harris, Hutaff-Lee, Koelemay, Dinkins, & Kirkwood, 2011; MacAllister, Nakhutina, Bender, Karantzoulis, & Carlson, 2009; Nagle, Everhart, Durham, McCammon, & Walker, 2006; Rienstra, Spaan, & Schmand, 2010; Schneider, Kirk, & Mahone, 2014).

In recent years, researchers have begun using simulation studies to examine whether children can malinger head injuries. Nagle and colleagues (2006) administered the TOMM and Hopkins Verbal Learning Test (HVLT) to children age 6-12. The HVLT is a brief test of verbal learning and memory. Each child was administered the HVLT twice and the examiners counterbalanced the order in which participants were told to try their best or to respond as if they had a head injury. The examiners did not provide any additional instructions on how to perform and there is no indication that the examiners instructed the participants to attempt to evade detection. For both groups, children were only administered the TOMM during the second testing session and both groups performed equally on this measure (M = 49.67 and M = 49.75 respectively), suggesting that they were unable or unwilling to intentionally reduce their performance on this test. This is not surprising as children have very limited knowledge of the symptoms associated with head injuries and it may be difficult for them to generate a reasonable strategy (Beardmore, Tate, & Liddle, 1999). For the HVLT, children who were instructed

to fake a head injury on the first trial scored equally on both trials, however, children who were instructed to fake a head injury on the second trial were able to significantly reduce their test score. One may hypothesize that children were only able to perform poorly on the second trial of the HVLT because they needed exposure to the test in order to develop a strategy to perform poorly. Alternatively, it is possible that children wanted to demonstrate their ability and could only feign an injury after being given the chance to do their best. DeRight and Carone (2015) note that the instructions given to children on how to feign a deficit was ambiguous and children may have tried their best because that is what they would expect children with brain injuries to do. Regardless, proper administration procedures dictate that these tests are only valid on the first administration, and under these conditions children were unable to perform poorly.

In a similar study, children age 6-11 were instructed either to try their best or to follow instructions to perform poorly on neuropsychological testing (Blaskewitz et al., 2008). The children who were instructed to perform poorly were provided a scenario in which a wizard was coming to their school to select children for his wizard school. Participants were told that they should get some questions wrong because the wizard did not want to recruit children who were smarter than him, but not so many questions wrong that the wizard knows they are faking. The PVTs included the TOMM, Medical Symptom Validity Test (MSVT; Green, 2003), and Fifteen Item Test (FIT; Rey, 1958). The tests of cognitive ability included the Trail Making Test, Digit Span (forward and backward), Digit Symbol, and Coloured Progressive Matrices. Comparing test performance across groups, the children in the simulating group performed worse than the best effort group on every test. All of the children in the best effort condition passed all three PVTs. With the exception of one child in the simulating group who performed well on every test, the rest of the children failed at least one of the PVTs and two children failed all of the PVTs. Sixty-eight percent of the children in the simulating group failed the TOMM and 90% failed the MSVT (i.e., they were detected as performing suboptimally). The authors note that fewer children may have failed the TOMM than expected because it was the last test to be administered for all children and poor attention may have impacted performance or adherence to the instructions. Despite this, the findings are contrary to those by Nagle et al. (2006) and suggest that children do *not* need prior exposure to the test to perform worse on cognitive tests and PVTs. An important limitation to this study is the use of broad age ranges that span a period of

rapid cognitive development, including the development of many abilities thought to be required to withhold one's best performance (such as the executive functions) (Gombos, 2006). With these rapid changes, it is not known whether there are developmental differences within the sample between children who could and could not withhold their best performance. Moreover, the result that only 68% of the children in the malingering condition failed the TOMM suggests that not all children followed the instructions. Lastly, it is not known whether these results would still be valid under standardized testing conditions or if the coaching was done further in advance of the testing session.

Gunn, Batchelor, and Jones (2010) also conducted a simulation study with children age 6-11. Participants were instructed to either try their best or to simulate memory impairments. Those in the simulation group were provided with specific instructions to follow, including, "Your goal is to try and produce the most severely impaired performance you can." In this study, the first two tests administered were the measures of performance, thus the instructions given to the children were fresh in their memory. The test order was not counterbalanced. Of the children instructed to try their best, only one failed the TOMM. Of those instructed to simulate memory impairment, 38/40 failed the TOMM. Although the results show that children can deliberately fail the TOMM when given specific instructions to get questions wrong, this scenario deviated substantially from a clinical situation (i.e. to feign impairment without being detected and embedding the PVTs within the test battery) and the results may not be generalizable.

In another simulation study, Green and colleagues (2012) reported the results of a Brazilian sample of children age 6-10 who were either asked to try their best or feign memory impairment. Of the children instructed to try their best, 98% passed the MSVT. Of the children instructed to feign memory impairment, all of them failed the MSVT (Green et al., 2012).

In addition to simulation studies, researchers have examined the failure rates on PVTs among samples of examinees who were eligible to apply for compensation due to their psychiatric injury; thus, base rates of malingering were suspected to be higher than in samples that were not eligible for compensation. Kirkwood & Kirk (2010) examined the scores of 1,973 children between the ages of 8-17 being assessed for a TBI. They reported that 17% of the sample failed the MSVT, of which 2% were believed to be false positives. Two percent (n = 3) were also thought to be false negatives as the children passed the MSVT but put forth variable performance on other tests. Similarly, Gidley-

Larson et al. (2015) found that 16% of children age 8-17 referred for mTBI assessments failed the MSVT. Chafetz et al. (2007) examined MSVT scores among children age 6-16 who were referred for assessments to determine social security disability benefits in the US and found that 37% failed the MSVT.

These failure rates are in sharp contrast to those of children with low motivation to malinger. Studies of children with ADHD (Harrison, Flaro, & Armstrong, 2015), moderate to severe brain damage/dysfunction (Carone, 2008), and Fetal Alcohol Spectrum Disorder (Gidley-Larson et al., 2015) all showed that the MSVT is sufficiently easy that only 5% of children failed the test. Carone (2014) described a case of a 9-year-old girl with severe congenital bilateral brain tissue loss, epilepsy, extremely low IQ, developmental delays, academic difficulties, and severely impaired adaptive functioning. Despite her cognitive status, she still passed the MSVT at almost perfect levels. The author suggests that less impaired examinees who fail the MSVT should be highly suspect of performing suboptimally.

At present, the literature on which children, and under what conditions, children can malinger is equivocal. There is some evidence that children do have the ability to deliberately perform poorly on cognitive testing, however, under what circumstances is unclear. The limited research in this area leaves open the question of whether children can malinger under ecologically valid conditions. That is, whether children can be coached to perform suboptimally prior to the testing session by someone other than the examiner, and whether children can deliberately perform poorly when the examiner uses standardized testing procedures (i.e., standardized instruction from the test manual and a single administration).

The research on detecting deception has overwhelmingly shown that the vast majority of adults cannot differentiate truthful from deceptive statements. As a result, those who need to detect deception have turned to using structured methods, such as specific neuropsychological tests that may alert evaluators that the examinee may be performing in a way that does not reflect their true ability. These tests were originally developed for use with adults in which secondary gain is possible and have now become part of routine practice in this context. There is emerging evidence that these tests can be used reliably with children as well. However, given a clinician's potential reliance on scores derived from these tests and the potential negative consequences to the

individual of obtaining a false positive result, there is a need for more research in this area.

The present study examines whether structured tests to detect suboptimal performance can be used with children, but before a child can deliberately feign an impairment on testing, they must have an understanding of the concept of lying and be able to behave in a manner that is consistent with their lie. In the following section, I will review the emergence of lying ability in children.

1.4. Development of Lying Ability in Children

Researchers have examined the emergence of lying in children as young as 2-3 years old (Evans & Lee, 2013). An adult may be motivated to lie for a variety of reasons, such as to obtain a benefit, avoid punishment, or for social-emotional reasons (e.g., avoidance of embarrassment or fear of reprisal). While children may lie for these reasons too, they tend to be less motivated to lie for financial gain because money is an abstract concept to them (Ceci & Bruck, 1993). In cases where children may be lying for financial gain, it is possible that they have been coached and/or encouraged to feign or exaggerate symptoms of a deficit by their parents or other adult with something to gain (Slick, Tan, Sherman, & Strauss, 2010). In the limited literature, this has been referred to as malingering by proxy (e.g., Lu & Boone, 2002).

To examine whether children have the ability to lie, one must consider their developmental capabilities (Rogers, 1997). As malingering is a complex task, it likely draws on many skills, such as having sufficiently developed theory of mind, controlling semantic leakage, second-order belief understanding, and sufficiently developed executive functions, in particular inhibitory control. There are likely many other processes that are involved, however, these have received the most attention in the literature.

One of the first requisite abilities to develop is theory of mind (ToM; Talwar & Lee, 2008; Williams, Moore, Crossman, & Talwar, 2016). This refers to the understanding that other people have beliefs, thoughts, and intentions that are different from one's own. This is necessary for successful deception since the purpose of lying is to create a false belief in the mind of another person (Talwar & Lee, 2008). Researchers

have shown that ToM begins to develop in the preschool years and by age 3 children understand that they can create a false belief by misleading others (Bussey, 1992; Chandler, Fritz, & Halla, 1989; Evans & Lee, 2013; Polak & Harris, 1999; Talwar & Lee, 2008; Wellman, 2014). Much of the research on children's lie-telling behaviour has focused on the emergence of lying in the preschool years (Evans & Lee, 2013; Polak & Harris, 1999; Talwar & Lee, 2002a; Talwar & Lee, 2002b; Talwar, Murphy, & Lee, 2007; Wellman, 2014). In a classic study, researchers told 3-year-old children that they were not allowed to peek at a toy when the researcher left the room (termed the *temptation resistance paradigm*). Invariably, almost all of the children peeked at the toy and 38% lied about peeking (Lewis, Stanger, & Sullivan, 1989). Replication studies have shown similar rates of verbal deception in children (e.g., Talwar & Lee, 2002a) with higher rates of lying among children age 4-7 than those age 3 (Talwar & Lee, 2002a; Polak & Harris, 1999). The results from the temptation-resistance paradigm indicate that simple deceptive ability, that is to falsely answer a "yes" or "no" question, develops at a very early age.

It has been noted that deception requires considerable skill beyond simply saying "yes" or "no", such as maintaining a lie and controlling one's verbal and non-verbal behaviour (Talwar, Gordon, & Lee, 2007). Researchers hypothesize that making a simple denial (e.g., "Did you take a peek at the toy?" –"No") only requires basic ToM (i.e., to represent a belief that is different from the truth); however, to maintain a lie children must know how to respond to subsequent questioning in a way that is consistent with their initial lie (Polak & Harris, 1999; Talwar & Lee, 2008). For example, assume a child is asked not to peek at a toy inside a box but they do because they cannot resist. If the child is then asked whether he or she peeked, it would be fairly easy to say "no." However, if the child is asked subsequent questions about the contents of the box, they would have to infer what response they should provide given their initial denial about peeking (i.e., fail to report the identity of the toy in the box). Being able to maintain a lie and respond consistently to subsequent questions is called *semantic leakage control* (Talwar & Lee, 2002a).

Talwar and Lee (2002a) examined the age at which children can control their semantic leakage. Using the temptation resistance paradigm, they asked follow-up questions to children age 3-7 years old who initially lied about having peeked at a toy. When children between the ages of 3 and 5 were asked the identity of the toy, they

blurted out the name of it. About half of the 6 and 7 year olds were able to feign ignorance of the toy's identity. Talwar, Gordon, and Lee (2007) used the same paradigm to elicit spontaneous lies from children age 6 to 11 years old. They asked follow-up questions and found that children's ability to remain consistent in their verbal statements increased with children's age. This ability was also significantly correlated with children's scores on a test of their *second order belief understanding* (SOBU). The authors suggest that children must be able to represent second-order mental states to lie consistently and control their semantic leakage (Talwar & Lee, 2002a). SOBU is best explained by an example: Suppose Jack and Jill are given a cookie to share. While Jill is playing outside, Jack hides the cookie in the cabinet. Unbeknownst to Jack, Jill was watching him through the window. If the child is able to represent second order mental states, they will understand that Jack does not know that Jill knows where he hid the cookie. They will also understand that Jack does not know where Jill will look for the cookie because Jack does not know that Jill saw him hide the cookie. If the child has acquired SOBU, when asked where Jack will think Jill will look for the cookie, the child will respond that he or she does not know. If a child has not acquired SOBU, the child will say that Jack thinks Jill will look for the cookie in the hiding place. It seems likely that the ability to successfully malinger on cognitive testing requires a minimum level of SOBU, as children must understand that the examiner holds a different set of beliefs in their mind (i.e., that the child is impaired), based on the information that has been made available to them thus far. Accordingly, children must maintain their lies during tests and consistently across tests to perpetuate their lie.

1.5. Executive Functioning.

Among the many cognitive processes that are implicated in deception, executive functions (EF) are said to play a primary role (Evans & Lee, 2011; Gombos, 2006; Williams et al., 2016). EFs are a set of interrelated cognitive processes that are required for goal-directed, non-automatic behaviour (Müller & Kerns, 2015). These cognitive processes begin to develop in early childhood and continue to develop into adolescence and adulthood (Müller & Kerns, 2015; Zelazo & Muller, 2002). Welsh, Pennington, and Groisser (1991) found that the EFs develop in a step-wise pattern and that different EFs

mature at different times. Similarly, imaging studies have shown that changes in EFs correspond to growth spurts in the frontal lobes and that developments in EFs are followed by periods of relative stability (Anderson, Anderson, Northam, Jacobs, & Catroppa, 2001; Müller & Kerns, 2015). Among researchers who study EF, there is debate as to which cognitive processes should be considered members of the category (Chan, Shum, Toulopoulou, & Chen, 2008). Some researchers argue that only "core processes," inhibitory control, set shifting, and updating/monitoring are true EFs (Miyake, Friedman, Emerson, Witzki, & Howerter, 2000). Other scholars consider a more extensive list of EFs that includes planning, organization, and working memory (Anderson, Anderson, Jacobs, & Smith, 2008). Of interest to the current series of studies and common to almost every model is the inclusion of inhibitory control as an EF (Carelli, Forman, & Mäntylä, 2008; Miyake et al., 2000; Müller & Kerns, 2015; Welsh & Pennington, 1988; Zelazo, Carter, Reznick, & Frye, 1997).

Inhibitory control is postulated to play a central role in successful deception (Carlson & Moses, 2001; Carlson, Moses, & Hix, 1998; Gombos, 2006; Williams et al., 2016). Inhibitory control refers to the ability to suppress dominant, automatic, or preponent responses in favor of more goal-appropriate ones (van der Sluis, de Jong, & van der Leij, 2007). This process plays a primary role in deception as the production of a lie requires the suppression of the truth (Gombos, 2006). Carlson and colleagues (1998) examined whether young children's difficulty with deception is attributed to a conceptual deficit or a lack of inhibitory control. In their study, children age 3-4 years old were able to deceive under conditions that required low inhibitory control (misleading pictorial cues or arrows) but were unable under conditions of high-inhibitory control (deceptive pointing). Furthermore, differences were unrelated to an understanding of false beliefs, which suggests that inhibitory control plays a role in producing simple deception. As the deceptive task becomes more difficult, understanding of false beliefs may play a more important role as the child needs to maintain their deception across tasks and time. Other researchers have examined the relationship between inhibitory control and lie-telling behaviour among very young children. Talwar and Lee (2008) used the temptation resistance paradigm to elicit spontaneous lies from children age 3-8. In their study, children who lied about having peeked at a toy scored higher on measures of inhibitory control than children who told the truth. While these studies lend support to the relationship between inhibitory control and lying, lying on neuropsychological tests is

more complicated and cognitively demanding than a single and simple lie about having peeked at a toy.

Petersen and colleagues (Petersen, Hoyniak, McQuillan, Bates, & Staples, 2016) conducted a large-scale review of studies measuring the development of inhibitory control from infancy through later childhood. They noted the problem of heterotypic continuity (the idea that the same construct may present differently at different ages), which highlights a limitation to measuring inhibitory control through the lifespan. This has an impact on the construct validity of various tests which purport to measure inhibitory control, as some tests may be more appropriate at certain ages than others. Among the studies they reviewed, the authors noted a rapid growth spurt in inhibitory control using age-appropriate tests among children age 2-4 (although inhibitory control as a specific executive function is not well-differentiated from the other executive functions at this age). The authors note that there is another growth spurt between the ages of 7-8, which is followed by a period of relatively gradual development. Other studies which extend beyond this age range (e.g., Macdonald, Beauchamp, Crigan, & Anderson, 2014) suggest a "leveling off" of the development of inhibitory control after age 8, with only modest gains through adolescence.

Variations of the Stroop test (Stroop, 1935) have been used for years to study inhibitory control in adults (MacLeod, 1991). Although there are various versions of the test, they all generally include the same format. In the classic version, the first trial includes an array of colour words printed in different colour inks and the examinee must read the words as quickly as possible. In the second trial, the examinee is again presented with an array of colour word printed in different colour inks, but the examinee must name the colour of the ink rather than reading the word. Different versions of the test are scored in different ways, but most generally score either the number of words read correctly within a given time frame, or the time it takes for the examinee to complete the task. Some scoring systems incorporate the number of errors when calculating the total scores. As the Stroop test relies on automatic word recognition, studies have found that it cannot be used reliably in children younger than seven years of age (Comalli, Wapner, & Werner, 1962; Wright, Waterman, Prescott, & Murdoch-Eaton, 2003).

While other tests exist to measure inhibitory control for children under age 7 (e.g., the Whisper Task; Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996),

they are generally limited to use within research studies and are not used in clinical assessment as they lack adequate empirical basis and psychometric properties. Studies have found that the interference effect elicited by the Stroop test is maximal in seven year olds and starts to decrease in later years (Comalli et al., 1962; Guttentag & Haith, 1978). Similarly, performance on the Stroop has been shown to be related to age with the strongest interference effects among grade 2 students (i.e., age 7; Dash & Dash, 1982). Finally, a study examining the development of executive functioning in 6 to 12 year olds found that children showed notable improvements on the Knock & Tap subtest of the NEPSY (Developmental NEuroPSYchological Assessment; which measures inhibitory control among other executive functions) between the ages of 6 and 7, after which only small improvements in inhibitory control were noted (Klenberg, Korkman, & Lahti-Nuuttila, 2001).

An important limitation to the measure of EFs is the task impurity problem (Rabbitt, 1998). Because the EFs are measured with tasks that also involve non-executive cognitive abilities, such as verbal ability or motor speed, it is impossible to create a pure measure of an EF. For example, in measuring inhibition using a Stroop task, an examinee must communicate a response using their verbal ability, and their performance is related to their processing speed and working memory. In addition, cognitive tasks often utilize more than one EF, which makes it difficult to isolate any specific one. To address the task impurity problem, some researchers have used a control task that utilizes the same processes, except for the relevant EF. Scores can then be compared to separate out the contribution of the EF (van der Sluis et al., 2007). Alternatively, it has been suggested that researchers use multiple measures of the same EF to converge on a more pure measure of the process (van der Sluis et al., 2007). In the present study, this was addressed by using both control tasks and multiple measures.

From a young age, children have the ability to deceive in simple and unsophisticated ways, such as denying they peeked at a toy. This ability relies on several cognitive skills that begin to emerge in early childhood and continue to develop over time. One of the key differences in lie-telling behaviour between young children and older children is that older children are better able to maintain their lie in responding to follow-up questions. This skill relies on the child having developed theory of mind and second-order-belief-understanding. These cognitive skills allow the child to know and

maintain an awareness that the receiver of their lie has different knowledge than they do and the child must maintain their lie over time. It is also believed that inhibitory control plays a key role in children's ability to deceive as children must inhibit their dominant response and generate an alternative and plausible response.

1.6. The Current Research Studies

The current series of studies was designed to examine whether children have the ability to withhold their best effort on neuropsychological tests that are commonly used to assess children's cognitive functioning. I hypothesized that children who are instructed to withhold their best effort will obtain lower scores on measures of memory and processing speed than children instructed to try their best. I also examined whether this ability was related to inhibitory control using both verbal and non-verbal measures of inhibition. I hypothesized that inhibitory control would be negatively correlated with withholding, such that children with better developed inhibitory control will be able to withhold their best effort to a greater extent than those with poorly developed inhibitory control. As it is common in assessments of adults to include tests to detect performance validity, I was also interested in whether the same tests could accurately differentiate children who were withholding from those who were trying their best. I hypothesized that tests used to detect suboptimal performance with adults would differentiate between those withholding their best effort and those trying their best among children. The studies in this dissertation follow a progression from being internally valid, where the testing session was conducted in a way to facilitate children's ability to withhold, to a design which closely resembled a typical neuropsychological testing session, in which children were not reminded or given prompts to withhold their best effort. I also included one study to address a question that has been raised in the literature regarding children's need to have prior exposure to the testing materials in order to withhold their best effort. I hypothesized that prior exposure to the testing materials would not be necessary for children to withhold their best effort.

To estimate the minimum number of participants required for this series of studies, I conducted a priori power calculations using G*Power (Faul, Erdfelder, Lang, &

Buchner, 2009) for each test of a primary hypothesis. For the hypothesis that there would be significant between group differences based on instructions given, the analysis is a 2 (instructions) x 2 (grade) repeated measures MANOVA with two dependent variables. Assuming a medium effect size, 80% power, and $\alpha = .05$, I would need a total sample of at least 34 participants, or 17 per grade cohort. To estimate the total number of participants required to test the second hypothesis, that inhibitory control is related to children's ability to withhold, I estimated a medium effect size for a linear regression given $\alpha = .05$ and 80% power. This yielded a minimum of 15 participants to detect an effect, assuming one is present, with 80% power. Note, this analysis only includes participants in the withholding condition, so the minimum number of participants is only for that condition. Lastly, to estimate the number of participants needed to test the hypothesis that tests to detect suboptimal performance can successfully differentiate between those who are withholding from those who are trying their best, the analysis is a 2 (instructions) x 2 (grade cohort) MANOVA with the MSVT subscales as the dependent variables. Assuming a medium effect size, 80% power, and $\alpha = .05$, this analysis would require a minimum sample of 32 participants. Overall, the a priori power analyses indicate that to test each of the primary hypotheses, I would need approximately 15-20 participants per group.

Chapter 2. Study 1

2.1. Method

Participants. All of the schools that were contacted to participate in this research study were within the Catholic Independent Schools of Vancouver Archdiocese (CISVA) as they have been receptive to participating in psychological research in the past. The CISVA contains 39 elementary schools in the Greater Vancouver Area. Once general consent from the school district was obtained, I contacted the principals of schools that had not been recently contacted by other students in my lab. This was done to reduce the risk of overloading specific schools with requests to participate in research studies. The school principals who provided consent typically discussed the request with the teachers of their grade 4 and 6 classrooms prior to providing consent. Teachers of participating classrooms were given a $50 gift card to a local bookstore and all children from participating classrooms were given a pen or pencil. Parents who provided consent were also entered into a draw for one of three cash prizes for $50, $150, or $250 to acknowledge their role in preparing their child for the second testing session.

Overall, there were 135 children in this study that were recruited from grade four (n = 63) and grade six classrooms (n = 75). Eighteen children were excluded from the sample because they could not demonstrate their understanding of the instructions (n = 5), missed one of the testing sessions (n = 4), met exclusionary criteria (described below; n = 4), asked to return to their classroom during the testing session (n = 2), declined participating (n = 2), or were feeling unwell (n = 1). Additionally, the parents of three children did not read the storybook to their child and thus their data was not included. Of the remaining participants, there were 46 children from grade four classrooms (Mage = 9.90 years, SD = 0.30 years, 44% male) and 68 children from grade six classrooms (Mage = 11.92, SD = 0.29, 40% male).

Measures. ***Inhibitory control.*** Children completed two measures of inhibitory control: the Stroop test and the Inhibition subtest of the NEPSY-II. The Stroop test was originally developed in 1935 and remains one of the most widely used tests of attention and response inhibition (MacLeod, 1991). Of the many versions, only the Golden Child Version (Golden, 1978) has norms available for children as young as six years old (Strauss et al., 2006). The task has three components, each with a 45 second time limit. The first page contains a grid of 100 colour words (red, green, and blue) printed in black ink for the child to read. The second page contains 100 X's that are printed in red, green, or blue, and the participant's task is to name the colours. The third page contains colour-words like the first page that are printed in colours from the second page. For example, the word "green" may be printed in blue ink. The task is for the participant to name the colour of the ink rather than read the word. The principle behind the test is that participants take longer to name the colour of the ink of a word than simply naming the colour or reading the word because of the color-word interference effect. The interference score is calculated by subtracting the colour score from the colour-word score (I = CW - C) and then converting the scores to T-scores (Strauss et al., 2006). The interference T-score tables are reversed, such that higher T-scores indicate more interference. In this dissertation, interference scores are noted as $Stroop_I$.

Children were also administered the Inhibition subtest of the NEPSY-II (Korkman, Kirk, & Kemp, 2007). The task has roots in the original Stroop test and involves looking at a series of shapes and arrows and naming either the shape, direction of the arrow, or an alternate response depending on the color of the shape or arrow. The Inhibition subtest correlates highly (r = .59) with the Color-Word Interference test of the Delis-Kaplan Executive Function System (DKEFS; Korkman et al., 2007), which is another version of the Stroop test. Although the tests correlate highly, the advantage of adding the Inhibition subtest of the NEPSY-II is that it is primarily non-verbal, which helps reduce the influence of verbal skills on measuring inhibitory control.

Reading ability. To ensure sufficient reading ability to complete the tasks, children were administered the WIAT-II-A Reading subtest (Psychological Corporation, 2002). The test was scored using the Canadian normative data. An a priori cut-off of two standard deviations below the normative mean was set in order to screen out children without sufficient single word reading skills as these children are likely to show an

attenuated interference effect on the Stroop and may have difficulty with other verbal tasks. No children obtained scores below this threshold to be removed from the sample.

Cognitive ability. Children were administered the Information subtest of the Wechsler Intelligence Scale for Children – 4th Ed. (WISC-IV). This subtest requires children to answer general knowledge questions and is a commonly used test of intelligence with strong psychometric properties and normative data for children age 6-16 years old. The Information subtest correlates at r = .73 with the Full Scale IQ (FSIQ), which makes it an adequate estimate of overall cognitive functioning.

Verbal memory. Children were administered the Rey Auditory Verbal Learning Test (RAVLT) as a measure of verbal memory. The RAVLT is one of the oldest verbal memory tests and is sensitive to neurological impairment (Strauss et al., 2006). This test was chosen as it is one of the few verbal memory tests with alternate versions and has been normed for use with children. It was also selected as I wanted to protect the integrity of other commonly used measures in case a study participant needed to be assessed for clinical purposes. The added benefit of this test is that it can be used without paying a fee to the publisher. The RAVLT can be used with children as young as seven years old and has separate norms for boys and girls (Vakil, Blachstein, & Sheinman, 1998). There are significant practice effects on the RAVLT and thus alternate versions were developed that have been shown to produce comparable results (Geffen, Butterworth, & Geffen, 1994). The version of the test used in this study consists of a list of 15 nouns that are read aloud to the examinee at a rate of one word per second. There are five trials that are each followed by a free recall test. The words and order of the list are the same for each trial. One of the scores that can be yielded from this test is a total score for trials 1-5. The score is simply a sum of the number of words recalled across the five trials. This score is one of the most reliable single scores that are yielded from the test (Strauss et al., 2006). Throughout this dissertation, reference to RAVLT scores indicates the sum score from trials 1-5 and will be noted as $RAVLT_{t1\text{-}5}$.

Processing speed. Children were administered two measures of processing speed from the WISC-IV: Digit-Symbol Coding (DS-Coding) and Symbol Search (SS). These two subtests were used to calculate the Processing Speed Index (PSI) on the WISC-IV. In DS-Coding, the child copies geometric symbols that are paired with numbers from a legend into boxes that only contain numbers. The child must correctly transpose as many symbols as possible in 120 seconds. In SS, the child must decide

whether a target symbol is present in an array of other symbols. If the target is present, they must draw a line through the “yes” box and if the target is absent, they must draw a line through the “no” box. The child has 120 seconds to respond to as many items as possible.

Performance validity. Children were administered the MSVT (Green, 2004), which is a shorter version of the Word Memory Test (WMT) with fewer word pairs and a shorter delay before the retention trial. The MSVT was originally designed to be used with both adults and children as young as 7 years old. Like other PVTs, the MSVT has been shown to be insensitive to neurological damage (Green, 2004). The test is verbal and includes learning and recalling a series of word pairs. The test is similar to other verbal memory tests that include both an immediate and delayed retention trial. Due to the protected nature of the test, I will not be describing the appearance of the test in more detail here. The test yields five scores, immediate recognition (IR), delayed recognition (DR), paired associates (PA), free recall (FR), and a consistency (CNS) score between the IR and DR trials. Insufficient effort is suspected if the examinee scores below 90% on the IR, DR, or CNS indices. The PA and FR trials are more difficult than the other trials and serve as measures of memory that are also sensitive to performance validity. Thus, lower scores on these scales may reflect individual differences in memory and are not used to detect performance validity. The test manual does not offer any other set of cut-offs for particular subgroups, such as children, and none have been proposed in the literature as more appropriate for children than the existing cut-offs.

Design and procedure. This study was a 2 (age: grade 4 or 6) x 2 (instruction: withhold or best effort) x 2 (session: session 1 or 2) mixed model design with age and instructions as between-subject factors and session as the within-subjects factor. This age was chosen as children younger than seven (grade 2) are rarely assessed by neuropsychologists as most tests do not have norms for children younger than this age. Moreover, I was interested in whether individual differences in withholding could be explained by inhibitory control and the best test to measure inhibitory control (Stroop) requires reading automaticity which is not consistently developed in children until approximately grade 3. Another reason to select this age range relates to the developmental research previously described, which shows that inhibitory control goes through a period of development around this age and it is hoped that individual

differences would be captured by including children during this period of developmental change.

Participants were randomly assigned to a set of instructions (best effort or withhold best effort), with the provision of having an approximately equal number of males and females in each group. All children participated in both testing sessions. With approval from the SFU Department of Research Ethics, CISVA, school principals, and teachers, an information sheet and consent form was sent home to parents outlining the study and requesting their consent to allow their child to participate. Attached to the consent form was a brief health history questionnaire about their child. The questionnaire was used to screen for a history of brain injuries, neurological problems, learning disabilities, colour-blindness, and ADD/ADHD (see Appendix A for a copy of the questionnaire). Children who met exclusionary criteria were still tested although their data was not included in the study.

The children of parents who provided consent were called out of class individually and asked if they were interested in participating in our research study (see Appendix B for verbal assent script). If children provided assent, they were taken to a quiet room and engaged in a brief conversation to develop rapport. Once the child appeared to be comfortable, the examiner briefly described the testing session and began the test battery. Children were allowed to discontinue the testing session and return to their classroom at any time.

At the start of the first testing session, children assigned to either instruction group were told to try their best on all of the tests given to them. As incentive to try their best, children were told that if they try their best, they would earn a prize at the end of the testing session. The tests were administered in counterbalanced order, with the limitation of having the appropriate time delay between trials of the MSVT and RAVLT and not overlapping these tests to avoid memory interference from word lists (see Appendix C for order of test administration). The tests administered in the first testing session were the following: RAVLT, NEPSY-Inhibition, WISC-Coding, WISC-Symbol Search, WISC-Information, Stroop, and WIAT-Reading. The tests administered in the second testing session were the following: MSVT, WISC-Symbol Search, WISC-Coding, and the RAVLT (alternate version to the version used in session 1).

At the end of the first testing session, children were told that they did an “excellent job” and were given a pen or pencil as a prize. Six days after the first testing

session, parents were given an envelope containing a storybook and instructions to read the story to their child in the evening to prepare them for the second testing session the following day. The envelope also contained a form for parents to sign and confirm that they read the story to their child. Parents were informed that their returned confirmation form would serve as their ballot for the cash prize. This served as an incentive to read the story and return the form. Based on the child's assigned instruction condition, the child's parent received one of two storybooks to read to their child. The books were also matched to the child's sex to increase the child's affiliation with the character in the book.

Parents of children in the withholding condition received a book about a child who bumped his/her head in a car accident and went to see a psychologist (see Appendix D for a copy of the storybooks). The psychologist administered a battery of psychological tests and the child performed as if they had suffered a mild TBI by demonstrating impaired memory, slowed processing speed, and difficulty with challenging questions. On the instruction page following the story, parents were asked to instruct their child to perform like the child in the storybook on testing. This meant that the child should (1) perform tasks slower than they normally could, (2) pretend that they cannot remember as much as they normally could, and (3) provide the wrong answer to questions that seem difficult.

To maintain parallelism between the instruction conditions, parents of children in the best effort condition received a storybook about a child who was having difficulty in math and went to see a psychologist. The psychologist in the story administered a battery of tests and the child worked hard on all of the tasks. The psychologist told the child to keep working hard at school and that math will become easier for them with practice. On the instruction page following the story, parents were asked to instruct their child to try their best on all of the tasks in the testing session the following day in school.

Before children began the second testing session, the examiner confirmed that the child read the storybook with their parent(s) and the instruction page following the story. Children were then asked to recite the one or three instructions in order to confirm their memory of the instructions and remind them to apply those instructions during the testing session. If a child was unable to recall all of the instructions, the examiner taught the instructions to the child until they could recall them without prompts. The child was then asked to demonstrate their understanding of the instructions by watching the examiner respond to a set of questions and then evaluate whether the examiner

followed the instructions. For example, if the examiner for a child in the withholding condition performed slowly on a measure of processing speed, the child would be correct in responding that the examiner followed the instructions. With regard to demonstrating poor memory, the examiner explained and demonstrated that if they were asked to remember three words, but pretended they could only remember two, they would be following the instruction. The examiner also demonstrated remembering all of the words, which was an example of not following the instructions as the examiner had remembered the word list too well. I chose to have the child demonstrate their understanding by evaluating the examiner rather than applying the instructions themselves as I wanted to confirm their understanding of the instructions and not whether they could apply the instructions; the testing session itself was intended to assess whether the child could apply the instructions. If the child correctly identified when the examiner was doing the task correctly, the examiner stated, “Okay, now you know what you’re supposed to do. If you remember to follow these instructions when we get started, you’ll get a special prize at the end. Okay, let’s get started” and began the testing session. If the child did not correctly identify when the examiner was following the instructions, the examiner reviewed the instructions and reassessed the child’s understanding a second time. If the child continued to make mistakes in identifying when the examiner followed the instructions, they were tested but their data was not included in the study. As previously indicated, this included four participants. The order of test administration was counterbalanced in session two as it was in session one, but the test order was not the same between testing sessions. Once the testing session had started, the instructions were not repeated to the child. If the child asked what they were supposed to do during the testing session, the examiner simply replied, “I want you to follow the instructions as best as you can.”

After the testing session, children were asked to remember the one or three things they had to do during the testing session. The number of instructions they recalled was recorded on the testing form. Children were debriefed about the study (see Appendix B for debriefing script) and given a prize for doing an “excellent job.” After the last child from a classroom had participated, the remaining children in the classroom were also given pens or pencils.

Debriefing. After the testing session, children were reminded that they were playing a pretend game like many other children’s games. They were thanked for trying

their best to follow the instructions and informed that if they are tested again in the future, they should try their best to get as many questions correct as possible. Children were provided an opportunity to ask questions about their participation and returned to their classroom.

With regard to motivation, one may argue that children would not be as motivated to withhold their best effort as they would be in a real life situation. While the potential reward in this study is substantially different from the potential financial reward in a civil litigation case, children are unlikely to be motivated by financial rewards because they are intrinsically less interesting to children and the reward would go to the parent. Children tend to prefer material rewards, such as pens, toys, and stickers. In addition, the instructions to withhold were given by the child's parent and an examiner, both of whom are in positions of authority.

2.2. Results

All test scores were converted to scaled scores using the appropriate age-based norms. When available, Canadian norms were used, which included subtest scores from the WISC-IV and WIAT-II. As previously indicated, only the RAVLT has gender-based norms, which were used to generate scaled scores. For the Stroop test, Interference scores were first calculated based on the formula I=CW-C (Strauss, Sherman, & Spreen, 2006) where I = interference, C = Colour Score, and W = Word score, and then converted to T-scores, such that high T-scores indicate greater Interference. For the Inhibition subtest of the NEPSY-II, total scores were calculated from tables in the manual that account for the examinees' time to complete the task and their number of errors. Scores on the WISC-IV Digit Symbol subtest and Symbol Search subtest were combined into a Processing Speed Index (PSI) using tables from the Canadian manual.

To check for possible errors in data entry, the range and distribution of scores for all variables were examined and confirmed to be accurate according to the original data records. Reading ability was assessed using the WIAT-II Word Reading subtest to rule out children who scored greater than two standard deviations below the mean. No participants met this criteria, and the mean scaled score was 11.02 (*SD* = 2.41) among fourth graders and 10.87 (*SD* = 2.99) among sixth graders.

The results of Study 1 are divided into three sections that address the following questions: (1) Are children who are instructed to withhold their best effort able to do so on the $RAVLT_{t1-5}$ and PSI? Does this change as a function of grade? (2) Can children's ability to withhold be explained by individual differences in inhibitory control? (3) Can the MSVT be used to detect children who are withholding their best effort?

Before I began the analyses, I examined all of the variables to assess whether they fell within normal limits as compared to the normative sample. All of the means were within one standard deviation of the normative sample mean except for scores on the RAVLT. See Table 1 for means and standard deviations for all variables reported by grade and instructions given. With regard to the low mean on the $RAVLT_{t1-5}$, I examined the distribution of scores for outliers using QQ plots and observed that the distributions were continuous. As the $RAVLT_{t1-5}$ score is a within-subject variable and I am not comparing the scores to a normative sample, the low mean does not pose a significant problem in this dataset.

Table 1. Mean (SD) scores for the $RAVLT_{t1-5}$ (Time 1 & 2), PSI (Time 1 & 2), Information, $Stroop_I$, and NEPSY-Inhibition tests as a function of grade and instruction given

		RAVLT: T_1	RAVLT: T_2	PSI: T_1	PSI: T_2	Info	$Stroop_I$	NEPSY: Inhib
Gr. 4	Best Effort (n = 20)	41.66 (12.78)	40.67 (13.56)	104.35 (11.14)	114.10 (10.92)	10.50 (2.46)	52.10 (12.14)	12.40 (1.82)
	Withholding (n = 26)	38.61 (11.53)	32.27 (14.40)	107.73 (15.50)	100.23 (21.06)	11.42 (2.34)	54.73 (10.11)	12.04 (2.84)
Gr. 6	Best Effort (n = 32)	33.76 (11.79)	39.30 (13.90)	105.62 (20.37)	113.31 (19.95)	10.56 (2.64)	49.41 (8.32)	11.50 (3.67)
	Withholding (n = 36)	35.92 (14.49)	28.05 (15.06)	104.58 (16.97)	92.86 (23.71)	11.14 (3.29)	49.17 (7.91)	11.39 (3.71)

With regard to alternate versions of the RAVLT, there were no significant differences between versions of the test at session one when both instruction groups were told to try their best, $t(112) = 0.09$, $p = .93$. To examine whether there were differences in versions of the test at session two, I conducted a 2 (instruction: withhold or

best effort) x 2 (RAVLT version) ANOVA on session two $RAVLT_{t1-5}$ scores and found no significant interaction between instructions and version of the test [$F(1, 3) = 1.01$, $p = 0.32$]. As such, the alternate versions of the RAVLT are confirmed to be equivalent in this study.

After the second testing session, participants were asked to recall the instructions they were asked to follow. The proportions of how many instructions participants in the withholding condition were able to remember are presented in Table 2. As all of the participants could remember at least one of the instructions and only a small proportion could only recall one instruction, no participants were excluded from the sample on these grounds. To examine whether there were differences between the number of instructions recalled between grade cohorts, I conducted a one-way ANOVA on number of instructions recalled. I only included participants in the withholding condition as there was no variability among children in the control condition. The difference in number of instructions recalled between grade cohorts was not significant, $F(1, 60) = 0.25$, $p = .62$.

Table 2. Number of instructions participants in the withholding condition could recall after the second testing session expressed as percentages separated by grade cohort

		0 Instructions	1 Instruction	2 Instructions	3 Instructions
Grade 4	$n = 26$	0%	8%	35%	58%
Grade 6	$n = 36$	0%	8%	42%	50%

Question 1: Are children who are instructed to withhold their best effort able to do so on the $RAVLT_{t1-5}$ and PSI? Does this change as a function of grade?

To examine whether the dependent variables (DVs; $RAVLT_{t1-5}$ and PSI) were related, I conducted correlations between them at session one and session two. At session one, the DVs were not significantly related ($r = .08$, $p = .20$), however, they were related at session two ($r = .38$, $p < .001$). As the DVs were unexpectedly not significantly related at session one, I visually inspected a scatterplot of the data for outliers and the data appeared to cluster. I also examined the data for outliers on the basis of scores

falling beyond three standard deviations from the mean and there were no outliers. On this basis, I continued to use the data as collected.

As this is a repeated measures design, it is of particular relevance to examine whether the change scores on each DV are related (i.e., $RAVLT_{t1\text{-}5\ Change\ Score}$ = Session $2_{RAVLT\ t1\text{-}5}$ – Session $1_{RAVLT\ t1\text{-}5}$). The $RAVLT_{t1\text{-}5}$ change score was significantly related to the PSI change score [r = .42, p < .001), thus a MANOVA is the appropriate analysis for the data.

A 2 (instruction) x 2 (time) x 2 (grade) mixed factorial MANOVA on $RAVLT_{t1\text{-}5}$ and PSI scores was conducted with instruction and grade as between-subjects factors and time as a within-subjects factor. I first assessed whether there were any differences by grade to see whether the grades could be collapsed. The main effect of grade was not significant [Wilks' Λ = .98, F(2, 111) = 1.83, p = .17, η^2_p = .03], nor was the three-way interaction between grade, time, and instructions [Wilks' Λ = .98, F(2, 111) = 0.97, p = .38, η^2_p = .02], or the two-way interactions between grade and instruction [Wilks' Λ = .99, F(2, 111) = 0.44, p = .65, η^2_p = .01] or grade and time [Wilks' Λ = .98, F(2, 111) = 1.33, p = .27, η^2_p = .02]. The non-significant effects were not surprising given that scores on the DVs had been scaled by age. To see whether participants in the withholding group performed differently than the control group across sessions, I examined the interaction between time and instruction. The result was significant [Wilks' Λ = .73, F(2, 111) = 20.74, p < .001, η^2_p = .27], indicating that the groups performed differently on the DVs across sessions. The univariate tests indicated that there was an effect of instruction and time for both the $RAVLT_{t1\text{-}5}$ [F(1, 110) = 10.61, p = .001, η^2_p = .09] and $PSI_{t2\text{-}t1}$ [F(1, 110) = 36.51, p < .001, η^2_p = .25)]. At the first session, the groups performed equally on the $RAVLT_{t1\text{-}5}$ [t(112) = -0.10, p = .92, d = -0.02, 95% CI of the Mean Difference (-5.10, 4.60)] and PSI [t(112) = -0.24, p = .81, d = -0.05, 95% CI of the Mean Difference (-7.02, 5.48). At the second session, the differences between the conditions emerged as children in the withholding condition obtained scores that were significantly lower than participants in the control condition on both the $RAVLT_{t1\text{-}5}$ [t(112) = 3.72, p < .001, d = 0.70, 95% CI of the Mean Difference (4.68, 15.34) and PSI [t(112) = 4.63, p < .001, d = 0.88, 95% CI of the Mean Difference (10.10, 25.23). Readers can refer back to Table 1 for the specific group means and standard deviations on the $RAVLT_{t1\text{-}5}$ and PSI as a function of instruction.

Question 2: Can children's ability to withhold be explained by individual differences in inhibitory control?

The second set of analyses examined whether inhibitory control can explain individual differences in children's ability to withhold. The mean and standard deviation scores on the $Stroop_I$ test and NEPSY Inhibition subtest were previously reported in Table 1. To examine the contribution of inhibitory control, above and beyond individual variability in cognitive ability, I used scores on the WISC-IV Information subtest as a covariate in these analyses as a rough estimate of crystallized intelligence (means and standard deviation scores are reported in Table 1). The correlation between WISC-IV Information scores and the WISC-IV Full Scale IQ score is .73 (Wechsler et al., 2003).

For these analyses, I used unstandardized inhibitory control scores as I was interested in absolute differences in inhibitory control across ages rather than relative scores within the child's age cohort. If I used standardized scores by age cohort, a child in grade 4 with a score of 100 would have equal inhibitory control to a child in grade 6 with the same score. The unstandardized means and standard deviations of $Stroop_I$ and NEPSY-Inhibition scores are reported in Table 3.

Table 3. Unstandardized means and standard deviations for participants in the withholding condition on the $Stroop_I$ and NEPSY-Inhibition tests as a function of grade

		$Stroop_I$	NEPSY: Inhibition
Grade 4	*n* = 26	-23.50 (10.77)	66.73 (16.96)
Grade 6	*n* = 36	-22.08 (7.64)	53.42 (11.02)

The participants in this analysis are only those children in the withholding condition, as I was only interested in the relationship between inhibitory control and withholding among those who were instructed to withhold. For each child, I calculated a change score as a measure of the degree to which they withheld. Change scores were calculated for both the $RAVLT_{t1\text{-}5}$ and PSI such that $RAVLT_{t1\text{-}5\ change\ score} = RAVLT_{t1\text{-}5\ Time\ 2} - RAVLTt_{1\text{-}5\ Time\ 1}$. Thus, a change score of -20 would indicate more withholding than a change score of -5. Mean change scores are presented in Table 4.

Table 4. Mean (SD) change scores for participants in the withholding condition as a function of DV and grade

		Mean (SD) $RAVLT_{t1-5}$ Change Score*	Mean (SD) PSI Change Score**
Grade 4	*n* = 26	-6.34 (15.12)	-7.50 (17.24)
Grade 6	*n* = 36	-7.87 (17.65)	-11.72 (20.45)

* $RAVLT_{t1-5}$ scores are *T*-Scores with *M* = 50, *SD* = 10
** PSI scores are Standard Scores with *M* = 100, *SD* = 15

In order to see whether inhibitory control explained any of the variability in each dependent variable beyond individual differences accounted for by the proxy measure of cognitive ability (WISC-IV Information), I constructed regression models for each DV with children's unstandardized scores on the Information subtest of the WISC-IV in the first block, grade cohort in the second block, unstandardized scores on the two measures of inhibitory control ($Stroop_I$ and NEPSY-Inhibition) in the third block, and the interaction between grade cohort and unstandardized inhibitory control scores in the fourth block.

Among children in the withholding condition, scores on the WISC-Information subtest did not predict scores on either the $PSI_{change\ score}$ [$R^2 < 0.01$, $F(1,60) = 0.14$, $p = .71$] or the $RAVLT_{t1-5\ change\ score}$ [$R^2 < 0.01$, $F(1,60) = 0.05$, $p = .82$]. Since the proxy measure of general cognition did not contribute to withholding related change scores, it was dropped from the full regression to simplify the model. The participants' grade cohort did not significantly explain any variability on the $PSI_{change\ score}$ [$R^2 = 0.01$, $F(1,60) = 0.73$, $p = .40$] or the $RAVLT_{t1-5\ change\ score}$ [$R^2 < 0.01$, $F(1,60) = 0.13$, $p = .72$]. There was no significant relationship between inhibitory control and either the $PSI_{change\ score}$ [$R^2_{change} = 0.04$, F(2,58) = 1.26, p = .29] or $RAVLT_{t1-5\ change\ score}$,[$R^2_{change} = .01$, $F(2,58) = 0.32$, $p = .73$], nor was the interaction between grade cohort and inhibitory control for either the $PSI_{change\ score}$ [$R^2_{change} = .02$, $F(2,56) = 0.70$, $p = .50$] or the $RAVLT_{t1-5\ change\ score}$ [$R^2_{change} = .01$, $F(2,56) = 0.42$, $p = .68$].

I further examined whether the relationship between change scores (on the PSI and $RAVLT_{t1-5}$) and inhibitory control scores may be better explained as a quadratic relationship as it is possible that children with high inhibitory control may be better at withholding an appropriate number of responses than children with poor inhibitory control who withhold too few or too many responses. As there were no differences in the

previous model between grade cohorts, they were collapsed in this regression. The relationship between change scores on the $RAVLT_{t1-5}$ and inhibitory control scores, however, was not better explained as a quadratic relationship, R^2 = .02, F(2, 59) = .69, p = .51, nor was there a significant quadratic relationship between PSI change scores and the measures of inhibitory control R^2 = .01, F(2, 59) = .33, p = .72.

As supplemental analyses, I was interested in what other factors might explain why some children were able to withhold their best effort while others could not. To examine this, I repeated the regression model with other variables in place of the inhibitory control scores. In particular, I was interested in whether memory was related to withholding. The two variables I had available that related to this were $RAVLT_{t1-5}$ scores at session one when participants were trying their best, and number of instructions recalled after session two. I conducted separate regression models for each of these variables as it would violate the assumption of independence to use session one $RAVLT_{t1-5}$ scores to predict $RAVLT_{t1-5}$ change scores. As there were no differences by grade and WISC-Information scores were unrelated to the DVs, I collapsed across grade and did not include WISC-Information scores in the first block.

Children's memory of the instructions after the second testing session significantly predicted change scores on the PSI, R^2 = .10, F(1,60) = 6.95, p = .01, β = -.32, and change scores on the $RAVLT_{t1-5}$, R^2 = .13, F(1,60) = 8.91, p < .01, β = -.36. Thus, children who remembered more of the instructions at the end of the second testing session obtained greater decreases in scores from session one to session two. Raw scores on the $RAVLT_{t1-5}$ at session one were unrelated to change scores on the PSI, R^2 = .05, F(1,60) = 3.46, p = .07.

Question 3: Can the MSVT be used to detect children who are withholding their best effort?

To investigate this question, I first looked at the MSVT in terms of assignment to withholding and control (best effort) conditions. Next, since the MSVT was designed to detect whether an individual actually withheld, rather than whether they were instructed to withhold, I categorized children as likely withholding or not based on their performance on the PSI and $RAVLT_{t1-5}$ and then reassessed whether the MSVT detected those who were likely withholding their best effort.

The mean and standard deviations for scores on the MSVT are presented in Table 5 as a function of instructions given and grade.

Table 5. Mean (SD) MSVT Scores Across Grade and Instruction Group

			IR	DR	CNS	PA	FR
Grade 4	Control	*n* = 20	99.25 (2.45)	99.75 (1.11)	98.50 (3.29)	99.50 (2.24)	72.25 (10.06)
	Withholding	*n* = 26	79.04 (20.20)	81.15 (22.11)	78.27 (20.88)	80.38 (24.41)	53.65 (14.94)
Grade 6	Control	*n* = 32	99.22 (2.24)	99.38 (1.68)	98.91 (2.45)	97.81 (5.53)	73.44 (15.58)
	Withholding	*n* = 36	80.83 (17.26)	82.08 (20.12)	78.75 (19.32)	82.50 (17.63)	57.08 (17.90)

The correlations between MSVT subscales are presented in Table 6. As there was a ceiling effect on the IR, DR, CNS, and PA scales for children in the control condition, correlations are only provided for children in the withholding condition as a function of grade.

Table 6. Correlations between subscales of the MSVT among participants in the withholding condition as a function of grade

			IR	DR	CNS	PA	FR
Grade 4	*n* = 26	IR	1				
		DR	.90**	1			
		CO	.72**	.69**	1		
		PA	.91**	.86**	.68**	1	
		FR	.65**	.61**	.56*	.74**	1
Grade 6	*n* = 36	IR	1				
		DR	.82**	1			
		CO	.82**	.77**	1		
		PA	.72**	.75**	.76**	1	
		FR	.62**	.69**	.61**	.73**	1

* Significant at the $p < .01$ level

** Significant at the $p < .001$ level

As the MSVT subscales were strongly related (see Table 6), I considered using a 2 (instruction) x 2 (grade) MANOVA on all five MSVT subscale scores with both factors as between-subjects variables. However, I noted that the magnitude of the difference between the variances of the withholding and control groups was quite large and all of the subscales failed Levene's homogeneity of variance tests (all p's < .001). Given the large disparity between the variances in the control and withholding groups for most of the MSVT subscales, I examined the distributional properties for each variable. Among fourth and sixth grade participants in the control conditions, scores on the IR, DR, CNS, and PA scales were all negatively skewed (ranges from -2.08 to -4.47) and leptokurtic (ranges from 3.18 to 20.00). Visually, the data for participants in the control condition showed a ceiling effect and although unimodal, was not normally distributed (Shapiro-Wilk for IR, DR, CNS and PA scales all had p's < 0.001). Thus, compared to acceptable limits in the literature for skewness and kurtosis of ±2 (Trochim & Donnelly, 2006; Field, 2000 & 2009; Gravetter & Wallnau, 2014), the distributions deviated far from normality.

Among participants in the withholding condition, scores on the IR, DR, CNS, and PA scales for both grade four and six participants were all negatively skewed, but the most extreme value was -1.17. The distributions for these four indices were all platykurtic with the highest kurtosis value being -1.2. Visually, the distributions for participants in the withholding condition were unimodal, but due to the sample sizes, the data was not entirely continuous. Scores at the ends of the tails were not sufficiently distant that they would be considered outliers (i.e. greater than 3 standard deviations from the mean), but the distributions had greater spread than was observed in the control conditions. In terms of normality of the distributions, the IR, DR, CNS, and PA scales all had Shapiro-Wilk *p* values that were less than or equal to .01.

Due to violating the assumptions for a MANOVA, specifically that the distributions were not normally distributed and the variances were not homogeneous, as well as the size of the cells not buffering against moderate violations of the assumptions (Lumey, Diehr, Emerson & Chen, 2002), I conducted the following analyses as non-parametric tests (Tabachnick B. G., Fidell L. S., 2001).

Given that the control and withholding groups were independent and there were multiple levels of the DV, the appropriate non-parametric test was the Kruskal-Wallis H test. The results showed no significant differences between the grade cohorts for any of the MSVT subscales [$\chi^2(1)$ ranged from .01 to .77 with all p's > .05]. Scores between the grade cohorts were thus collapsed for the analysis on the effect of instruction. Across all five MSVT subscales, there were significant between group differences based on instructions given [all $\chi^2(1)$ were greater than 28.77 and all p values were less than .001 (see Table 7 for specific results). In terms of direction of the difference, the group of participants who were instructed to put forth their best effort scored significantly higher on all five MSVT subscales than the group who were instructed to withhold their best effort.

Table 7. Statistical tests for the effect of instructions on the MSVT subscales

MSVT Scale	Chi Square	**Cohen's *d***
IR	$\chi^2(1) = 49.54, p < .001$	1.46
DR	$\chi^2(1) = 37.02, p < .001$	1.21
CO	$\chi^2(1) = 40.52, p < .001$	1.43
PA	$\chi^2(1) = 32.99, p < .001$	1.13
FR	$\chi^2(1) = 28.77, p < .001$	1.14

n = 114

To examine whether the degree of withholding on the $RAVLT_{t1-5}$ and PSI was related to scores on the MSVT subscales, I correlated change scores on the $RAVLT_{t1-5}$ and PSI with scores on the MSVT. As there are multiple correlations, I corrected for possible type 1 error by setting alpha to .01. As before, I only included participants in the withholding condition, as there was insufficient variance among participants in the control condition on the MSVT. Among fourth grade children assigned to the withholding condition, there were significant correlations between all scales on the MSVT with change scores on the PSI (see Table 8 for correlations), but no significant correlations between scores on the MSVT with change scores on the $RAVLT_{t1-5}$. Among sixth grade participants, there was an identical pattern of non-significant correlations between change scores on the $RAVLT_{t1-5}$with scores on the MSVT, but significant correlations

between change scores on the PSI and all scales of the MSVT. As the pattern of results between grade cohorts was the same, the specific correlations are reported in Table 8 collapsed across grade.

Table 8. Correlations between change scores on the PSI and $RAVLT_{t1-5}$ with scores on the MSVT among children in the withholding

	IR	DR	CO	PA	FR
PSI	.51**	.62**	.49**	.55**	.48**
$RAVLT_{t1-5}$	.18	.20	.10	.26*	.26*

n = 62

* indicates significance of $p < .01$

** indicates significance of $p < .001$

Before assessing the classification accuracy of the MSVT, I checked the assumptions for calculating sensitivity and specificity; that the sample is representative of the greater population, each data point is independent, and there is a known "gold standard" that is error free (Lu, Fang, Tian & Jin, 2003). In this instance, the sample is assumed to be representative and the data points are independent, however, I had to assume that assigned condition is the "gold standard" against which the accuracy of the MSVT is evaluated. This is a limitation that is discussed further in the Discussion section. I first categorized each participant as having passed or failed the MSVT using the existing cut-offs from the test publisher's manual. As a reminder, failure on the MSVT is classified by at least one score that is at or below 85% on the IR, DR, or CNS scales. I note that there may be another cut-off that is more appropriate with children, however, the cut-offs that I use here are the only ones proposed by the test author that are used clinically, are used elsewhere in the literature, and the sample is not large enough to calculate what other cut-off may optimize the classification accuracy of the test. Consistent with not finding a significant effect of grade, the proportion of grade four and six participants who were classified as true positives, true negatives, false positives, and false negatives, were almost identical and thus collapsed across age for subsequent analyses; see Table 9.

Table 9. Proportion of participants that were classified as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) by the MSVT

	Withhold Group	Control Group
Fail MSVT	TP = 58.0% (n = 36/62)	FP = 0% (n = 0)
Pass MSVT	FN = 42.0% (n = 26/62)	TN = 100% (n = 52/52)

With assigned instruction as the criterion, the sensitivity of the MSVT was .58 and the specificity was 1.0. This suggests that the test performs very well at eliminating false positives (i.e., identifying someone as putting forth suboptimal effort when they had actually tried their best), but only detects 58% percent of those assigned to the withholding condition as putting forth suboptimal effort. It is worth noting that since the proportions of participants across grades were almost identical, the test performs equally across ages in this sample, using the test publisher's recommended cut-offs. Furthermore, older participants were not better at evading detection than younger children. The positive likelihood ratio cannot be mathematically computed as the formula would require a division by zero, but it is essentially perfect as a positive test result has a perfect likelihood of coming from a participant who was assigned to the withholding condition. The negative likelihood ratio is 2.38, which suggests that a pass on the MSVT is 2.38 times as likely to have come from someone in the control condition than the withholding condition. The positive predictive power (PPP) is the probability that an individual with a positive test result (detected on the MSVT) was in the withholding condition. The PPP of the MSVT in this sample was 1.0, indicating that 100% of participants identified as withholding were from the withholding condition. The negative predictive power (NPP) is the probability that an individual not identified by a test (passed the MSVT) was in the control condition. The NPP of the MSVT was .67, indicating that 67% of participants identified by the MSVT as putting forth optimal effort were assigned to the best effort condition.

As the MSVT is not designed to detect whether a participant was *instructed* to withhold their best effort, but rather to detect *actual* suboptimal effort, I was also

interested in examining the classification accuracy of the MSVT based on whether participants were likely withholding their best effort. I set the criteria for "likely withholding" based on participants obtaining a -1SD negative change score on either or both the PSI and $RAVLT_{t1-5}$. I set the cut-off based on the degree of withholding observed in other similar studies (Blaskewitz et al., 2008; 1.2-1.8 SDs below the mean) and took the further step of accounting for the error associated with each test (PSI and RAVLT) by calculating a Reliable Change Index[1] (RCI) score for each participant (Jacobsen & Truax, 1991). The proportion of participants who met the criteria for likely withholding did not overlap entirely with assigned condition (see Table 10). Due to the small sample sizes, the groups were collapsed across grade.

Table 10. Proportions of participants in each instruction group who obtained negative RCI scores greater than one standard deviation on either the PSI or $RAVLT_{t1-5}$ as a function of instruction given.

	Control	Withholding
$RAVLT_{t1-5}$	7.7% (*n* = 4/52)	43.5% (*n* = 27/62)
PSI	13.5% (*n* = 7/52)	50.0% (*n* = 31/62)

Using performance as the criterion to assess classification accuracy, however, did not allow for the accurate calculation of sensitivity, specificity, PPP, NPP, or likelihood ratios as the base rates in some cells were too low (see Podell, DeFina, Barrett, McCullen & Goldberg, 2003). When this occurs, the metrics become increasingly unreliable and are less likely to be replicable across samples. In addition, the same issue of not having a "gold standard" introduces error into evaluating the accuracy of the MSVT to differentiate between groups, and thus there is further uncertainty that the metrics generated to assess the accuracy of the MSVT to differentiate between groups using this criteria would be accurate. For this reason, they are not reported here.

[1] The formula for the calculation is as follows: $(S_2 – S_1)/SE_D$, where S_1 is the examinees initial test score and S_2 is their retest score on the same measure. SE_D is the Standard Error of the Difference and is an index of measurement error. It is calculated as $SE_D = \sqrt{[2\text{-}(SEM)^2]}$.

Additional analyses: Further examination of participants instructed to withhold who passed the PVT tests.

As follow-up analyses, I was interested in further examining the characteristics of the group of participants who were assigned to the withholding condition but passed the MSVT. While this group would typically be termed "successful malingerers," I was interested in whether they altered their performance or in fact tried their best. The group consisted of 11 fourth grade participants (42% of grade 4 withholding group) and 15 sixth grade participants (42% of grade 6 withholding group). The mean change score on the $RAVLT_{t1-5}$ was -10.65 (*SD* = 12.94) among fourth graders and -3.14 (*SD* = 15.16) among sixth graders. On the PSI, there was a mean difference of -0.09 (*SD* = 12.71) among fourth graders and -3.14 (*SD* = 12.24) among sixth graders.

Within this subsample, change scores on the PSI and $RAVLT_{t1-5}$ were not significantly correlated (r = .21, p = .31), thus two separate ANOVA's or t-tests were appropriate. To examine whether there were differences between children in grade four and grade six, I used two separate t-tests with grade as the between-subjects factor and change scores on the PSI and $RAVLT_{t1-5}$ as the DVs. There were no differences between grades on either the PSI [$t(24) = 0.50$, p = .63] or $RAVLT_{t1-5}$ [$t(24) = 1.33$, p = .20].

Sixty-two percent of participants assigned to the withholding condition obtained negative RCI scores on at least one measure that was greater than one standard deviation, and 15% (n = 4) obtained negative RCI scores on both measures greater than one standard deviation. Thus, 38% of the participants in the withholding group put forth their best effort or did not withhold enough to make a meaningful difference in their performance. It is interesting that three out of four participants who obtained negative RCI change scores greater than one standard deviation on both measures also obtained at least one negative RCI change score greater than three standard deviations, which would likely indicate faking or severe impairment. If I suppose that participants who obtain negative RCI change scores greater than three standard deviations on either the PSI or $RAVLT_{t1-5}$ would likely be detected as obviously faking, I am left with 18% of children in the withholding condition who passed the MSVT and obtained negative RCI change scores on at least one measure between one and three standard deviations. While this may seem high, it is important to consider the base rate of participants in the

best-effort condition who also passed the MSVT and obtained at least one negative RCI change score between one and three standard deviations. Among participants in the best-effort condition, 19% passed the MSVT and obtained at least one negative RCI change score between one and three standard deviations. Thus, the groups are in fact comparable in terms of the proportion of participants who obtained negative RCI change scores between one and three standard deviations. This suggests that the group of participants who were instructed to withhold their best effort but were not detected by the MSVT performed comparably to the group of participants who tried their best. In other words, this group did not evade detection, but rather did not follow the instructions to withhold their best effort.

2.3. Discussion

For Study 1, I hypothesized that children asked to withhold their best effort during the second testing session would obtain lower scores on the $RAVLT_{t1\text{-}5}$ and PSI during that session than children asked to put forth their best effort during both sessions. The results showed that participants in the withholding condition obtained significantly lower scores during the second session than participants in the best effort condition on both the $RAVLT_{t1\text{-}5}$ and PSI. Interestingly, the magnitude of effect was larger on the PSI than the $RAVLT_{t1\text{-}5}$. There were no significant differences between the grade cohorts, which is not surprising since scores were scaled by age. These results are consistent with previous studies that show that young children have the capacity to withhold their best effort under certain conditions (see Kirkwood, 2015b for a review).

The second hypothesis was that among children asked to withhold their best effort, individual differences in children's ability to withhold would be related to individual differences in inhibitory control. Participants completed two measures of inhibitory control, the Stroop task and the Inhibition task from the NEPSY. Both measures are similar, however, the Stroop task is verbal and the Inhibition task is non-verbal. For both the PSI and $RAVLT_{t1\text{-}5}$, change scores were calculated (i.e., $RAVLT_{t1\text{-}5\ T2} - RAVLT_{t1\text{-}5\ T1} = RAVLT_{t1\text{-}5\ \text{Change Score}}$) to obtain a single score of the degree of withholding. I accounted for individual variability in cognitive ability by putting WISC-IV Information scores as a proxy for general intelligence into the first block of a regression model and then both

inhibitory control scores into the second block. The measures of inhibitory control did not significantly explain any variability either collectively or independently in withholding on the PSI or $RAVLT_{t1\text{-}5}$ across both grade cohorts. I followed up by examining whether the relationship may not be linear, but it was not better explained as a quadratic function. Although there was sufficient power to detect a medium effect size, it is possible that the relationship is subtler and there was inadequate power to detect a smaller effect.

The third hypothesis in Study 1 was that the MSVT could be used to detect children who were performing suboptimally. There was a significant main effect of instructions on MSVT scores as a whole and across all five individual MSVT subscales such that children in the withholding group had lower scores than children in the best effort group. The effect sizes across all five MSVT subscales were large (Cohen's *d* ranged from 1.13-1.46), indicating large between-group differences. While there was a main effect of instruction on MSVT scores, when group membership was the criterion, the test obtained excellent specificity but mediocre sensitivity. Among participants in the best effort condition, the MSVT correctly classified 100% of the participants as belonging to that condition. Thus, the test did not incorrectly detect any participants as putting forth suboptimal effort. Among participants in the withholding group, however, the MSVT correctly classified 58% as belonging to the withholding condition and misclassified 42% as belonging to the best effort condition. This suggests three possible explanations: (1) some participants recognized the PVT and evaded detection, (2) the recommended cut-offs for the test achieved only slightly better than chance identification rates of those who were withholding, or alternatively, (3) not all of the participants who were instructed to withhold actually withheld. The data supports the third explanation for a few reasons: among participants who were instructed to withhold but were not detected by the MSVT, there was an equal number of participants in the withholding and best effort conditions who obtained negative RCI change scores between -1 *SD* and -3 *SD* below the group mean on either the RAVLT or PSI. Thus, not all of the participants in the withholding group withheld their best effort because as a group, they obtained comparable scores to the best effort group. Furthermore, the range and degree of variance in scores among participants in the withholding group suggests that some of them performed comparably to the control condition on the performance based tests and at the ceiling of the MSVT. Lastly, it seems unlikely that participants were able to successfully identify the MSVT as a PVT although I did not ask them to identify which test they believed to be a PVT.

In summation, the results from Study 1 show that children have the ability to withhold their best effort on measures of memory and processing speed under certain conditions. It has, however, been argued in the literature that children need prior exposure to the testing materials in order to withhold their best effort (Nagle et al., 2006). Since the repeated measures design of Study 1 allowed participants to put forth their best effort in session one prior to withholding their best effort during session two, I conducted a second study to examine whether the results are replicable when participants do not have prior exposure to the testing materials before being asked to withhold.

Chapter 3. Study 2

In a previous study by Nagle et al. (2006), children were asked to malinger either during the first or second testing session. They found that only children who were asked to malinger on the second session were able to perform suboptimally. The authors concluded that children need prior exposure to the tests in order to malinger successfully. To address this possibility, I repeated the first study with only one testing session. Due to time constraints at the end of the school year, all of the children in Study 2 were asked to withhold. None of the children in Study 2 had participated in Study 1. As children were sampled from the same school district and within two months of the previous study, I compared data from Study 2 to the data from the control condition from Study 1. Scores from Study 1 were taken from the first administration of the $RAVLT_{t1\text{-}5}$ and PSI to eliminate practice effects. Table 11 presents a comparison of the age and sex of withholding participants from Study 2 to the comparison group from Study 1 (note, the data represents the groups after participants were removed because they did not meet criteria for inclusion). Due to the methodological limitations of this approach, which are discussed below, the results from Study 2 should be interpreted with caution.

Table 11. Comparison of Study 2 participants to comparison group

	Grade 4 M_{age} in Months (*SD*)	**Grade 6 M_{Age} in Months (*SD*)**	**Percentage male**
Comparison Group	118.65 (3.02) (n = 20)	143.14 (3.59) (n = 32)	40.9%
Study 2 Withholding Group	120.04 (3.85) (n = 20)	141.79 (3.41) (n = 24)	37.2%

3.1. Method

Participants. There were 47 children in this study that were recruited from grade four (n = 23) and grade six (n = 24) classrooms. Three children were excluded from the sample because two of them could not demonstrate understanding of the instructions and one was feeling unwell. Of the remaining participants, there were 20 children from grade 4 classrooms (M_{age} = 10.00, SD = 0.32) and 24 children from grade 6 classrooms (M_{age} = 11.82, SD = 0.28). As previously indicated, permission to recruit students to participate in this study was obtained from the CISVA (school board), school principals, and teachers. Teachers of participating classrooms were given a $50 gift card to a local bookstore and all children from participating classrooms were given a pen or pencil. Parents who provided consent were also entered into a draw for one of three cash prizes for $50, $150, or $250 to acknowledge their role in preparing their child for the testing session.

Measures. The measures used in Study 2 were the same as those used in Study 1 with the addition of the TOMM (Tombaugh, 1997). The TOMM was added to Study 2 because the testing session was shorter than Study 1 and I wanted to make the most of the time I had with the study participants. Moreover, clinicians rarely rely on a single PVT and adding the TOMM would allow me to make comparisons between the two PVTs. The TOMM is one of the most common measures of performance validity among neuropsychologists (Sharland & Gfeller, 2007). It appears to measure memory, however, it is insensitive to memory deficits associated with mild to moderate traumatic brain injury, aphasia, cognitive impairment, and dementia (Hill et al., 2003; Tombaugh, 1997). There is some evidence that it misclassifies those with severe cognitive impairment and moderate to severe dementia as malingering (Strauss et al., 2006).

The TOMM has two learning trials and a delayed retention trial. The test is non-verbal and has the appearance of a memory test. As the test is protected, I will not be describing the appearance of the test in any more detail. The maximum score is 50/50 and any score less than 45 on trial 2 should raise concern that the individual may be putting forth suboptimal effort. Although the TOMM has been researched more extensively with adults than children, studies have shown that it can be used reliably with children as young as five. In one study of healthy children putting forth optimal effort, the

mean scores for 5-year-olds on trial 2 was 49.71/50.00 (*SD* = 0.48) (Constantinou & McCaffrey, 2003). In another study, children age 6-12 obtained mean scores ranging from 49.90-50.00/50.00 (Nagle et al., 2006). In a simulation study of children age 6-11 who were told to malinger, 68% of the children obtained scores that were below the cut-off (Blaskewitz et al., 2008).

Design and procedure. This study was a 2 (age: grade 4 or 6) x 2 (instruction: withhold or best effort) factorial design with age and instructions as between-subjects factors. As previously indicated, all children in Study 2 were instructed to malinger. I used data from the control group's first testing session in Study 1 as the comparison group for Study 2 and thus did not use random assignment for instruction given. As such, readers should interpret the results from Study 2 with caution.

The recruitment and consent procedures were identical to those in Study 1. The day before the testing session, parents were given an envelope containing a storybook and instructions to read the storybook to their child in the evening to prepare them for their testing session the following day. As before, storybooks were matched to the child's sex to increase the child's affiliation with the character in the book. The envelope also contained a form for parents to sign and confirm that they read the story to their child. As an incentive to read the storybook and confirm having done so, parents were informed that their returned confirmation form would serve as their ballot for the cash prize. As in Study 1, the last page of the storybook contained instructions for the parents to teach their child to perform like the child in the storybook during the testing session in school. The storybook and three instructions for withholding were the same as those in Study 1.

Before children began the testing session, the examiner confirmed that the child read the storybook with their parent and the instruction page following the story. Children were asked to recall the instructions from the storybook to confirm their memory of the instructions and remind them to apply those instructions during the testing session. If a child was unable to recall all three of the instructions, the examiner taught the instructions to the child until they could recall them without prompts. As in Study 1, children were then asked to demonstrate their understanding of the instructions by evaluating whether the examiner was correctly following them. If the child did not correctly identify when the examiner was following the instructions, the examiner reviewed the instructions and reassessed the child's understanding a second time. If the child continued to make mistakes in identifying when the examiner followed the

instructions, they were tested but their data was not included in the study. Once the testing session had started, the instructions were not repeated to the child. If the child asked what they were supposed to do during the testing session, the examiner simply replied, “I want you to follow the instructions as best as you can.”

After the testing session, children were asked to remember the three things they had to do during the testing session and the number of correct responses was recorded. After the testing session, children were debriefed as they were in Study 1.

3.2. Results

As in Study 1, the range and distribution of scores for all variables were examined for possible errors in data entry and confirmed to be accurate according to the original data records. Scores were converted to standard scores using the appropriate norms as they were in Study 1. The following results section for Study 2 is divided into two sections that address the following questions: (1) Are children who are instructed to withhold their best effort able to do so, even without prior exposure to the testing materials? Does this change as a function of grade? (2) Can the TOMM and MSVT be used to detect children who are withholding their best effort?

Question 1: Are children who are instructed to withhold their best effort able to do so on the $RAVLT_{t1\text{-}5}$ and PSI, even without prior exposure to the testing materials? Does this change as a function of grade?

Mean scores on the RAVLT and PSI are presented in Table 12 as a function of grade and instruction given.

Table 12. Mean (SD) scores for the $RAVLT_{t1\text{-}5}$ and PSI for participants in the withholding and comparison groups as a function of grade and instruction given

			$RAVLT_{t1\text{-}5}$	**PSI**
Gr. 4	Comparison Group	(*n* = 20)	41.66 (12.78)	104.35 (11.14)

	Withholding Group	(*n* = 20)	38.94 (15.26)	92.20 (16.87)
Gr. 6	Comparison Group	(*n* = 32)	33.75 (11.79)	105.62 (20.37)
	Withholding Group	(*n* = 24)	35.67 (17.62)	101.00 (21.38)

To examine whether the DVs (i.e., $RAVLT_{t1\text{-}5}$ and PSI) were related, I correlated RAVLT and PSI scores for participants in the withholding group from Study 2 and found that the DVs were significantly correlated, (r = .21, p = .046). As such, a MANOVA was used to analyze the data.

A 2 (instruction: withhold or best effort) x 2 (grade) MANOVA on $RAVLT_{t1\text{-}5}$ and PSI scores was conducted with instructions and grade as between-subjects factors. I first assessed whether there were any differences by grade to see whether the grades could be collapsed. The interaction between grade and instructions was not significant [Wilks' Λ = .99, $F(2, 91) = 0.63$, p = .09, η^2_p = .01], however, the main effect of grade was significant [Wilks' Λ = .93, $F(2, 91) = 3.36$, p = .04, η^2_p = .07]. As there were significant differences between grade cohorts, subsequent analyses are separated by grade. I re-examined the correlations between the DVs by grade and found that the DVs were not significantly correlated among fourth graders (r = .14, p = .38) but they were significantly correlated among sixth graders (r = .29, p = .03). As such, the analyses on fourth grade participants are ANOVAs and the analyses on sixth grade participants are MANOVAs.

For fourth grade participants, I conducted two separate one-way ANOVAs on $RAVLT_{t1\text{-}5}$ scores and PSI scores with instructions as the between-subjects factor. The effect of instructions on $RAVLT_{t1\text{-}5}$ scores was not significant [$F(1, 38) = 0.37$, p = .55, d = .19, 95% CI of the Mean Difference (-6.29, 11.73)], however, it was significant for PSI scores [$F(1, 38) = 7.23$, p = .01, d = .85, 95% CI of the Mean Difference (3.00, 21.30)]. For sixth grade participants, I conducted a MANOVA on $RAVLT_{t1\text{-}5}$ and PSI scores with instructions as the between-subjects factor. Despite expectations based on developmental logic, the effect of instructions was not significant for sixth grade participants [Wilks' Λ = .97, $F(2, 53) = 0.63$, p = .54, η^2_p = .02]. While the effect of instructions was not significant at the group level, it is conceivable that some participants in the withholding group were indeed withholding their best effort and thus I decided to proceed with question 2 below.

Question 2: Can the TOMM and MSVT be used to detect children who are withholding their best effort?

The mean and standard deviations for scores on the MSVT are presented in Table 13 and the mean and standard deviations for scores on the TOMM are presented in Table 14. As before, there were large differences in variances between the MSVT scores in the withholding and comparison group (Levene's test of homogeneity *p*'s were all < .01, except for the FR scale). The differences in variances suggests that contrary to the lack of between instruction group differences reported above, there were differences in how the participants responded in the testing session based on the instructions given, albeit on different tests.

Table 13. Mean (SD) MSVT scores across grade for participants in the withholding group and comparison group

			IR	DR	CO	PA	FR
Grade 4	Comparison Group	(*n* = 20)	99.25 (2.45)	99.75 (1.11)	98.50 (3.29)	99.50 (2.24)	72.25 (10.06)
	Withholding Group	(*n* = 20)	83.25 (23.24)	88.75 (22.18)	84.50 (17.98)	88.50 (19.81)	64.25 (12.59)
Grade 6	Comparison Group	(*n* = 32)	99.22 (2.24)	99.38 (1.68)	98.91 (2.45)	97.81 (5.53)	73.44 (15.58)
	Withholding Group	(*n* = 24)	83.12 (20.37)	85.00 (23.55)	83.96 (18.65)	87.08 (23.31)	68.75 (20.01)

Table 14. Mean (SD) TOMM scores for participants in the withholding group as a function of grade

		Trial 1	Trial 2
Grade 4	(*n* = 20)	41.45 (8.72)	45.50 (11.40)
Grade 6	(*n* = 24)	41.79 (11.20)	43.17 (12.50)

Correlations between MSVT subscales are presented in Table 15. Since participants in Study 1 did not complete the TOMM, there is no comparison group to

which I can compare the results of participants who were instructed to withhold on the TOMM. Thus, the only psychometrics I can calculate regarding classification accuracy of the TOMM using assigned instructions as the criterion is the sensitivity.

Table 15. Correlations between subscales of the MSVT among participants in the withholding condition as a function of grade

		IR	DR	CO	PA	FR
Grade 4	IR	1				
	DR	.87**	1			
	CO	.80**	.51*	1		
	PA	.77**	.91**	.46*	1	
	FR	.47*	.69*	.15	.74**	1
Grade 6	IR	1				
	DR	.80**	1			
	CO	.88**	.70**	1		
	PA	.63*	.90**	.48*	1	
	FR	.55*	.65*	.37	.65*	1

* Significant at the $p < .01$ level

** Significant at the $p < .001$ level
$n = 44$

As the MSVT subscales were strongly related (see Table 15), I considered using a 2 (instruction: withhold or best effort) x 2 (grade) MANOVA on all five MSVT subscale scores with both factors as between-subjects variables. However, as I found in Study 1, the magnitude of the difference between the variances of the withholding and comparison groups was quite large. On further examination, the IR, DR CNS and PA subscales all failed Levene's test of homogeneity of variances at the $p \leq .001$ level, and the FR scale had a $p = .05$. Because the scores for the comparison group came from Study 1, I will only briefly repeat that the distributions had a negative skew, were

leptokurtic, and were not normally distributed (with the exception of the FR scale, which was normal). Among fourth graders in the withholding group, skewness ranged from -2.82 to -0.78, and kurtosis ranged from 0.43 to 8.48. For the IR, DR, CNS and PA subscales, the scores were not normally distributed (Shapiro-Wilk *p*'s ≤ .001). Visually, the distributions were unimodal and negatively skewed with scores towards the tails not continuous, although there was not sufficient distance that they would be considered outliers (i.e. > 3 standard deviations from the mean). Among sixth graders, skewness ranged from -0.78 to -2.02 and kurtosis ranged from -1.00 to 4.09. For the IR, DR, CNS, PA and FR subscales, the scores were not normally distributed (IR, DR, CO and PA subscales Shapiro-Wilk *p*'s < .001, FR subscale Shapiro-Wilk *p* = .03). Visually, the distributions were unimodal but skewed and had scores towards the tail that were not continuous, but were not sufficiently distant that they would be considered outliers.

Given that the assumptions for MANOVA were not met and the sample sizes were not sufficiently large to buffer against violations of the assumptions, the appropriate non-parametric test is the Kruskal-Wallis H test. The results showed no significant differences between the grade cohorts for any of the MSVT subscales [$\chi^2(1)$ ranged from .01 to 1.61 with all p's > .05]. Scores between the grade cohorts were thus collapsed for the analysis on the effect of instruction. Across all five MSVT subscales, there were significant between group differences based on instructions given [all $\chi^2(1)$ were greater than 3.85, with all p values less than .05 (see Table 16 for specific results)]. In terms of direction of the difference, the group of participants who were instructed to put forth their best effort scored significantly higher on all five MSVT subscales than the group who were instructed to withhold their best effort.

Table 16. Comparison of MSVT subscale scores between instruction groups

MSVT Scale	Chi Square	Cohen's *d*
IR	$\chi^2(1) = 29.78, p < .001$	1.05
DR	$\chi^2(1) = 18.03, p < .001$	0.79
CO	$\chi^2(1) = 28.65, p < .001$	1.12
PA	$\chi^2(1) = 9.60, p < .01$	0.69

FR	$\chi^2(1) = 3.85, p < .05$	0.41

n = 96

To assess the classification accuracy of the MSVT and TOMM based on instruction given, I first categorized each participant as having passed or failed each test according to the test publishers' user manuals. For the proportions of participants that were classified as true positives, true negatives, false positives, and false negatives on the MSVT, see Table 17. For the proportions of true positives and false negatives on the TOMM, see Table 18.

Table 17. Proportion of participants that were classified as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) by the MSVT

	Withhold Group	**Comparison Group**
Fail MSVT	TP = 59.5%	FP = 0%
Pass MSVT	FN = 40.5%	TN = 100%

n = 96

Table 18. Proportion of participants in the withholding group that were classified as true positives (TP) and false negatives (FN) by the TOMM

Fail TOMM	TP = 22.2%
Pass TOMM	FN = 77.8%

n = 44

As above, I checked the assumptions for calculating sensitivity and specificity and aside from not having a "gold standard", which is a limitation that is discussed later, the assumptions were met. When instruction given was used as the criterion, the sensitivity of the MSVT was .60 and the specificity was 1.0. Similar to Study 1, the MSVT performed very well at eliminating false positives, but only detected 60% of those assigned to the withholding condition as putting forth suboptimal effort. As before, the positive likelihood ratio cannot be computed but is essentially perfect and the negative likelihood ratio is 2.47. The PPP was 1.00 and the NPP 0.71.

No participants from the comparison group completed the TOMM and thus many of the psychometrics cannot be computed for the TOMM. The sensitivity, however, was .22.

As before, I was interested in the classification accuracy of the MSVT and TOMM according to whether participants actually withheld their best effort and not whether they were instructed to withhold their best effort. As before, I set the cut-off at -1 *SD* as I was interested in subtle withholding and previous studies found that children instructed to malinger obtained scores between 1.2 - 1.8 standard deviations below the mean (Blaskewitz et al., 2008). As participants only completed each measure once, I used their actual scores rather than their RCI scores. To assess which participants obtained scores lower than one standard deviation below the mean, I used the mean and standard deviations from the first session of Study 1 for both the PSI and $RAVLT_{t1-5}$ as the means in this sample were higher on the PSI and lower on the $RAVLT_{t1-5}$ than the normative sample. For participants in grades 4 and 6, I used the respective means for their cohort from Study 1. The proportions of participants in each instruction group who obtained scores below this cut-off are presented in Table 19 as a function of instruction group. Scores were again collapsed by grade as there was no main effect or interaction between the cohorts based on instructions given. It is important to note that both the $RAVLT_{t1-5}$ and PSI total scores are normally distributed (RAVLT: Shapiro-Wilk = .99, p = .36, PSI: Shapiro-Wilk = .99, p = .85), and thus I would expect a base rate of approximately 16% of the participants to obtain scores at or below one standard deviation below the mean on each test. For the proportions of participants that were classified by the MSVT as true positives, true negatives, false positives, and false negatives, see Table 20.

Table 19. Proportions of participants in each group who obtained scores lower than one standard deviation below the mean on either the PSI or $RAVLT_{t1-5}$ as a function of grade and instruction group.

	Comparison Group	**Withholding Group**
$RAVLT_{t1-5}$	13.5% (*n* = 7/52)	15.9% (*n* = 7/44)
PSI	15.4% (*n* = 8/52)	36.4% (*n* = 16/44)

Table 20. Proportion of participants that were classified by the MSVT as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as a function of obtaining scores on either or both measures above or below the -1 SD cut-off.

PSI +/ $RAVLT_{t1-5}$ Score	**MSVT Fail**	**MSVT Pass**
≤ -1 *SD* cut-off	TP = 15.6% (*n* = 15/96)	FN = 19.8% (*n* = 19/96)
> -1 *SD*	FP = 9.4% (*n* = 9/96)	TN = 54.2% (*n* = 52/96)

Using the -1 *SD* or greater cut-off as the criterion for withholding, however, did not allow for the accurate calculation of sensitivity, specificity, PPP, NPP, or likelihood ratios as the base rates in some cells were too low (see Podell, DeFina, Barrett, McCullen & Goldberg, 2003), even with scores collapsed by grade. Thus, the results would not be reliable and are not reported for that reason.

With regard to the TOMM, the sample only includes participants who were instructed to withhold. The proportions of participants that were correctly and incorrectly classified according to the cut-off of -1 *SD* on either or both measures are presented in Table 21.

Table 21. Proportion of participants that were classified by the TOMM as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as a function of obtaining scores on either or both measures above or below the -1 SD cut-off

	PSI +/ $RAVLT_{t1-5}$ Score	**TOMM Fail**	**TOMM Pass**
Grade 4	≤ -1 *SD* cut-off	TP = 15.0% (*n* = 3/20)	FN = 35.0% (*n* = 7/20)
	> -1 *SD*	FP = 0% (*n* = 0/20)	TN = 50.0% (*n* = 10/20)
Grade 6	≤ -1 *SD* cut-off	TP = 20.8% (*n* = 5/24)	FN = 16.7% (*n* = 4/24)
	> -1 *SD*	FP = 12.5% (*n* = 3/24)	TN = 50.0% (*n* = 12/24)

Using the cut-off of -1 *SD* or greater on either or both the $RAVLT_{t1-5}$ and PSI as the cut-off for likely withholding, the base rates were again too small to accurately

calculate the classification accuracy of the TOMM. One can see from the scores, however, that only a small number of participants with scores less than -1SD were detected by the TOMM and half of the participants with scores greater than -1SD passed the TOMM. The data suggests that a number of participants were classified as "likely withholding" based on the performance cut-offs but passed the TOMM, meaning that the test either did not detect them, or they were not withholding their best effort and their score below the cut-off was part of the normal distribution of scores that extends below -1SD.

Interestingly, there was agreement between the MSVT and TOMM in only 7 cases out of the group of 19 who failed at least one effort test. Among the cases where only one PVT detected suboptimal effort, it was always the MSVT that detected the participant and the participant had passed the TOMM.

Additional analyses: Further examination of participants instructed to withhold who passed the SOP tests.

Other studies have shown that failure on at least one PVT is associated with a decrease in performance on ability-based tests that are administered in the same testing session (Kirkwood, Yeates, Randolph, & Kirk, 2012). Applied to clinical practice, failure on a PVT suggests that low scores on ability tests should be interpreted with caution, as the examinee may not be putting forth optimal effort. To see whether the data from the present study supported this conclusion, I divided the sample of participants who were instructed to withhold their best effort into two groups based on whether they passed both PVTs or failed at least one PVT. This resulted in 18 participants in the group that passed both PVTs and 26 participants in the group that failed at least one PVT. For the means and standard deviations of scores on the $RAVLT_{t1-5}$ and PSI based on whether participants had passed both effort tests or failed at least one, see Table 22 below.

Table 22. Mean (SD) scores among participants instructed to withhold their best effort as a function of whether or not they passed both PVTs

		RAVLT	PSI
Grade 4	Passed Both Effort Tests (*n* = 7)	37.25 (15.01)	94.57 (16.68)
	Failed at Least One PVT (*n* = 13)	39.86 (15.93)	90.92 (17.51)

Grade 6	Passed Both Effort Tests (*n* = 11)	42.50 (7.61)	112.36 (19.32)
	Failed at Least One PVT (*n* = 13)	29.89 (21.62)	91.38 (18.61)

Among fourth graders in the withholding group (*n* = 20), scores on the PSI and $RAVLT_{t1\text{-}5}$ were significantly correlated (*r* = .49, *p* = .03), but among sixth graders (*n* = 24), they were not (*r* = .33, *p* = .12). Although the DVs were not significantly correlated among sixth graders, the magnitude of the correlation was moderate and the non-significant finding is likely attributable to the small sample sizes. As such, I ran the model as a single MANOVA so that I could also include grade as a between-subjects factor.

For the MANOVA, grade and whether a participant passed both PVTs or failed at least one were the between-subjects factors. Scores on the $RAVLT_{t1\text{-}5}$ and PSI were the dependent variables. The main effect of grade was not significant [Wilks' Λ = .92, $F(2, 39) = 1.77$, p =.18, η^2_p = .08], nor was the main effect of having passed/failed at least one PVT [Wilks' Λ = .89, $F(2, 39) = 2.36$, p =.11, η^2_p = .11] or the interaction between grade and having passed or failed at least one effort test [Wilks' Λ = .97, $F(2, 39) = 2.72$, p =.29, η^2_p = .08].

As before, I was interested in further examining the group of participants that were assigned to the withholding condition but were not detected by either the MSVT or TOMM. While the main effect of having passed both PVTs or failed at least one was not significant, I thought it may have been possible that some participants in the withholding group may have been withholding but their performance was masked by those who did not follow the instructions and rather tried their best. Of the 44 participants assigned to the withholding condition, there were seven fourth graders (35% of grade 4 withholding group) and eleven sixth graders (46% of grade 6 withholding group) who passed both PVTs. Among this group of participants in the withholding condition who were not detected by either PVT, only one fourth grade participant obtained a score lower than one standard deviation below the comparison group mean on both the $RAVLT_{t1\text{-}5}$ and PSI. There were no other participants in this group who obtained scores below the same threshold on the $RAVLT_{t1\text{-}5}$. There were three additional participants who obtained scores below the same threshold on the PSI; two in grade four and one in grade six. Collectively, two of these scores were between one and two standard deviations below the mean and two were between two and three standard deviations below the mean. In

other words, 4/18 or 22% of the participants in the withholding condition passed the MSVT and TOMM but obtained scores on at least one measure that was lower than one standard deviation below their comparison group mean.

To examine whether there were differences between children in grade four and grade six within the subsample of individuals who were assigned to the withholding condition but passed both PVT's, I used two separate one-way ANOVAs with grade as the between subjects factor and scores on the PSI and $RAVLT_{t1\text{-}5}$ as the DVs. There were no differences between grades on either the PSI [$F(1, 16) = 4.01$, $p = .06$] or $RAVLT_{t1\text{-}5}$ [$F(1, 16) = 0.98$, $p = .34$].

We can test the hypothesis that the proportion of participants who passed the PVTs and obtained at least one score lower than one standard deviation below the comparison group mean is the same as the proportion in the comparison group (H_0: $p_1 = p_2$) by calculating a z-score and corresponding p-value. The formula is as follows: $z = [(\hat{p}_1 - \hat{p}_2) - 0] / \sqrt{[\hat{p}\,(1-\hat{p})(1/n_1 + 1/n_2)]}$, where n are the respective samples of participants who passed the PVTs, $\hat{p}$ is the proportion of those who obtained a score on at least one measure below the aforementioned threshold over those who did not. In terms of the comparison group, I selected participants who passed the PVT and obtained a score on either measure that was lower than one standard deviation below the group mean. It is important to note that participants in Study 1 did not complete the TOMM and thus were selected on the basis of having passed the MSVT alone, however, I do not believe this is problematic as the pass rate on the TOMM among participants trying their best in other studies is 100% (e.g. Constantinou & McCaffrey, 2003). Among the 52 participants in the comparison group, all of them passed the MSVT. Of this group, 14 participants obtained a score on either the $RAVLT_{t1\text{-}5}$ or PSI that was lower than one standard deviation below the group mean. In testing whether this proportion of participants in the comparison group (14/52)[2] was the same as the proportion of participants who met the same criteria in the withholding group (4/18), I obtained a z-score of 0.42 ($p = .66$), and thus failed to reject the null. This suggests that as a group, the participants in the withholding condition who passed the PVTs did not simply evade detection, but rather they did not withhold their best effort, or withheld so minimally that it was not detected.

[2] I note that the proportion of participants in Study 1 who fell below the threshold of -1 *SD* was based on using RCI change scores, whereas in this study, the threshold is based on obtaining a score lower than -1 *SD* below the comparison group mean. This explains why 19% of the best effort participants in Study 1 met this criterion, whereas 27% met this criterion in Study 2.

In Study 1, I found that memory for the instructions was significantly related to change scores, such that participants who remembered more of the instructions after the testing session obtained greater negative change scores than those who remembered fewer instructions. In the current study, I do not have change scores or a baseline for each participant and thus individual differences introduces significant variance into this relationship. As I found in Study 1, memory for the instructions was negatively correlated with PSI scores (r = -.34, p = .03), which is consistent with the earlier reported finding of a main effect of instructions given (withhold vs. optimal) on PSI scores. Similarly, memory for the instructions was not significantly correlated with $RAVLT_{t1-5}$ scores (r = -.17, p = .28). This is particularly interesting because participants completed both the PSI and RAVLT as part of the same testing session and suggests that participants who remembered the instructions were better able to apply them to the PSI than RAVLT.

3.3. Discussion

To address the question in the literature of whether children need to have prior exposure to the specific tests in order to withhold their best effort (Nagle et al., 2006), I conducted a second study that only included a single testing session to see whether the results from Study 1 would be replicated. This partial replication would also bolster the findings from Study 1. This methodological change moved the study closer to a "real world" scenario where presumably the examinee would not have seen the specific tests before completing them. To keep other experimental variables the same as the first study, children were given the same instructions as they were in Study 1 and the examiner similarly ensured that participants understood the instructions before beginning the testing session. One additional difference is that with the additional time available during Study 2, participants were also administered the TOMM as a second PVT.

The first hypothesis was that children instructed to withhold their best effort would obtain lower scores on both the $RAVLT_{t1-5}$ and PSI than children who were instructed to try their best. As there were some significant differences in scores between the grade cohorts, scores were analyzed separately by grade. In addition, it is important to remind the reader that due to time constraints with the end of the school year, participant in Study 2 were only instructed to withhold their best effort and their scores are being

compared to participants from the first session of Study 1. Thus, as there was no random assignment, these results should be interpreted with caution. Among fourth grade students, participants who were asked to withhold their best effort obtained significantly lower scores on the PSI but not significantly different scores on the $RAVLT_{t1-5}$. Among sixth grade students, there were no significant differences between the groups on either the PSI or $RAVLT_{t1-5}$. Interestingly, the effect size for the significant difference on the PSI among fourth graders, which was the only significant result within this set of results, was large (d = .85).

While the only between-group difference based on instructions given was on the PSI among fourth graders, there were differences among sixth graders between those who passed both PVTs and those who failed at least one. Specifically, the sixth graders who failed at least one PVT obtained significantly lower scores on the PSI and somewhat lower scores on the $RAVLT_{t1-5}$ than those who passed both PVTs. This is consistent with most other studies which show that individuals who fail at least one PVT obtain lower scores on ability-based tests than those who pass PVTs. This finding supports the notion that clinicians should be wary of low scores on ability-based measures when the examinee has failed a PVT. This data further supports the notion that there was indeed an effect of instructions and that the experimental manipulation had an effect, albeit not a statistically significant one on the performance based tests. It appears that the effect was either too small to detect, or the group itself responded to the manipulation in a heterogeneous way and the scores from individuals who withheld were masked by those who did not.

The second hypothesis in this study was that the recommended cut-offs for adults on the MSVT and TOMM could also be used to differentiate children who were not putting forth their best effort from those who were. Because the TOMM was not administered to the comparison group in Study 1, there are limited statistics for the efficacy of this measure in this study. For both grade cohorts who were in the comparison group and instructed to try their best, the mean score on the MSVT for the three indices (IR, DR, and PA) used to make a decision about performance validity exceeded the recommended cut-offs and was near 100% accuracy. Among the groups instructed to withhold their best effort, the mean scores on the same three indices fell below the recommended cut-offs. In addition, it is noteworthy that the range of scores and degree of variance on the MSVT among participants in the withholding group was

much broader than participants in the comparison group. Typically, a large difference in the magnitude of the variance between conditions is considered a nuisance that potentially violates assumption testing (hence using a non-parametric test for related analyses). In this study, however, it is my view that the substantial increase in variance shows that the instructions given had an impact on performance.

Although the MSVT appears to function well at the group level using the recommended cut-offs, when individual participants were classified as having passed or failed the MSVT, the classification accuracy wavered. The test remained excellent at correctly identifying those who were instructed to put forth their best effort, but was only moderately accurate, and below acceptable standards, at correctly identifying those who were instructed to withhold their best effort. These results were comparable between the grade cohorts. On the TOMM, the classification accuracy of participants instructed to withhold their best effort was only moderate when using the recommended cut-offs.

As I was interested in the classification accuracy of the PVTs among those who actually withheld their best effort rather than those who were instructed to withhold, I set criteria for test performance on the PSI and $RAVLT_{t1-5}$ that likely indicated withholding based on the effect sizes of scores among participants in other studies who successfully withheld their best effort. Using this criterion for whether a participant was likely withholding, I again assessed the classification accuracy of the PVTs. It is important to note that in using a performance-based criterion, I expected at least some of the participants to fall below the cut-off based on the normal distribution of scores. More specifically, as the scores were normally distributed, I expected approximately 16% of the participants to obtain scores below the cut-off on each test. With this performance-based criterion, the classification accuracy of the MSVT was only mediocre at detecting those who withheld and rejecting those who did not. The classification accuracy of the TOMM was similarly unsatisfactory with the exception of the sensitivity of the test among fourth graders. If we accept that some participants who were instructed to withhold their best effort did not follow the instructions, it makes sense that the PVTs did not detect them and what appears to be mediocre classification accuracy actually reflects the finding that some children in this group put forth their best effort and were not detected for that reason.

I conducted a follow-up analysis of participants that were instructed to withhold their best effort but were not detected by either PVT. If participants in this group in fact

withheld their best effort, they could be considered “successful withholders” as they evaded detection. The other possibility is that this group did not withhold as instructed and actually put forth their best effort and thus were not detected because they were not performing suboptimally. There were a total of 18 participants that were instructed to withhold but passed both PVTs. Within this group, a small minority obtained scores that were lower than one standard deviation below the comparison group mean on the $RAVLT_{t1-5}$ or PSI, however, this proportion was not statistically different than the proportion of participants that obtained similarly low scores in the comparison group. Thus, the data suggests that these participants as a group were unlikely to be withholding and had tried their best or withheld so minimally that it was not detected.

Lastly, similar to the finding in Study 1 that memory for the instructions after the testing session was related to performance, I found that memory for the instructions was negatively correlated with PSI scores but not significantly correlated with $RAVLT_{t1-5}$ scores. This is consistent with the significant between group effect of instructions (withhold vs. optimal performance) on PSI scores noted above. As such, it appears that a critical factor in whether children can withhold their best effort is whether they have remembered the instructions until the end of the testing session.

Chapter 4. Study 3

As a follow-up to Study 2, I wanted to see how children would perform without a reminder of the instructions from the examiner immediately prior to the testing session. In this procedure, children were also unaware whether the examiner knew they had been instructed by their parents to withhold their best effort. As before, participants read the storybook with their parents the night before the testing session and were instructed by their parents to withhold their best effort, however, the examiner did not review the instructions or confirm that they knew how the child had been instructed to perform on the tests. Due to time constraints, I was not able to randomly assign participants to a best effort condition and all participants in Study 3 were instructed to withhold. As children were sampled from the same school district and within two months of the previous study, I compared the data from Study 3 to the data from the control condition from Study 1. Scores from Study 1 were taken from the first administration of the RAVLT and PSI to eliminate practice effects. A comparison of the age and sex of withholding participants from Study 3 to the comparison group from Study 1 is presented in Table 23. Due to the methodological limitations of this approach, the results from this study should be interpreted with caution.

Table 23. Comparison of Study 3 participants to comparison group

	Grade 4 M_{age} in Months (*SD*)	Grade 6 M_{Age} in Months (*SD*)	**Percentage male**
Comparison Group	118.65 (3.02) (n = 20)	143.14 (3.59) (n = 32)	34.6%
Study 3 Withholding Group	113.90 (3.42) (n = 20)	143.01 (3.39) (n = 28)	43.8%

4.1. Method

Participants. There were 50 children in this study who were recruited from grade four (n = 22) and grade six (n = 28) classrooms. The data from one child was removed from the sample because he asked to go back to his classroom part way through the testing session and another was excluded because she was not feeling well. Of the remaining participants, there were 20 children from grade 4 classrooms (M_{age} = 9.49, SD = 0.29) and 28 children from grade 6 classrooms (M_{age} = 11.92, SD = 0.28). As previously indicated, permission to recruit students to participate in this study was obtained from the CISVA (school board), school principals, and teachers. Teachers of participating classrooms were given a $50 gift card to a local bookstore and all children from participating classrooms were given a pen or pencil. Parents who provided consent were also entered into a draw for one of three cash prizes for $50, $150, or $250 to acknowledge their role in preparing their child for the testing session.

Measures. The measures used in Study 3 were the same as those used in Study 2. Readers can refer to the preceding sections for a description of the measures used.

Design and procedure. This study was a 2 (age: grade 4 or 6) x 2 (instruction: withhold or best effort) factorial design with age and instructions as between-subjects factors. As previously indicated, I used data from the control group's first testing session in Study 1 as the comparison group for Study 3 and thus did not use random assignment for instruction condition. As such, readers should interpret the results from Study 3 with caution.

The recruitment and consent procedures were identical to those in Studies 1 and 2. The day before the testing session, parents were given an envelope containing a storybook and instructions to read the storybook to their child in the evening to prepare their child for the testing session the following day. The storybooks used were the same as in Studies 1 and 2. As before, storybooks were matched to the child's sex to increase the child's affiliation with the character in the book. The envelope also contained a form for parents to sign and confirm that they read the story to their child. As an incentive to read the storybook and confirm having done so, parents were informed that their returned confirmation form would serve as their ballot for the cash prize. As in Studies 1

and 2, the last page of the storybook contained instructions for the parents to teach their child to perform like the child in the storybook during the testing session in school. The storybook and three instructions for withholding were the same as those in Studies 1 and 2.

Unlike in Studies 1 and 2, the examiner did not confirm with the child that they read the storybook with their parents the night before the testing session, nor did the examiner ask the child to demonstrate their understanding of the instructions prior to the testing session. The examiner read the standard testing instructions that ask the child to try their best. If the child asked what they were supposed to do during the testing session, the examiner simply replied, “I want you to follow the instructions as best as you can.” No indication was given about which set of instructions the child should follow.

After the testing session, the examiner explained that they knew the child had been instructed by their parent to follow a set of instructions and praised the child for doing a good job. Children were then asked to remember the three things they had to do during the testing session and the number of correct responses was recorded. After the testing session, children were debriefed as they were in Studies 1 and 2.

4.2. Results

As in Studies 1 and 2, the range and distribution of scores for all variables were examined for possible errors in data entry and confirmed to be accurate according to the original data records. Scores were converted to standard scores using the appropriate norms. The following results section is divided into two sections that address the following questions: (1) Are children who are instructed to withhold their best effort able to do so, even without prior exposure to the testing materials, without a reminder of the instructions prior to the testing session, and with an examiner who seemingly does not know the child has been instructed to withhold their best effort? Does this change as a function of grade? (2) Can the TOMM and MSVT be used to detect children who are withholding their best effort?

Question 1: Are children who are instructed to withhold their best effort able to do so on the $RAVLT_{t1-5}$ and PSI, even without prior

exposure to the testing materials, without a reminder of the instructions to withhold before the testing session, and with an examiner who seemingly does not know the child has been instructed to withhold? Does this change as a function of grade?

Mean scores on the $RAVLT_{t1-5}$ and PSI are presented in Table 24 as a function of grade and instruction given.

Table 24. Mean (SD) scores for the $RAVLT_{t1-5}$ and PSI for participants in the withholding group and the comparison group as a function of grade and instruction given

			$RAVLT_{t1-5}$	**PSI**
Gr. 4	Comparison Group	(*n* = 20)	41.66 (12.78)	104.35 (11.14)
	Withholding Group	(*n* = 20)	43.38 (15.39)	95.15 (18.60)
Gr. 6	Comparison Group	(*n* = 32)	33.75 (11.79)	105.62 (20.37)
	Withholding Group	(*n* = 28)	39.77 (16.50)	104.14 (18.42)

To examine whether the DVs (i.e., $RAVLT_{t1-5}$ and PSI) were related, I correlated $RAVLT_{t1-5}$ and PSI scores for participants in Study 3 and the comparison group from Study 1 and found that the DVs were not significantly correlated, (r = .17, p = .25). As such, ANOVAs are the appropriate analysis for the data.

I conducted separate ANOVAs for each DV. The first was a 2 (instruction: withhold or best effort) x 2 (grade) ANOVA on $RAVLT_{t1-5}$ scores with instruction given and grade as between-subjects factors. I first assessed whether there were any differences by grade to see whether the grades could be collapsed. The interaction between grade and instructions was not significant [$F(1, 96) = 0.55$, p = .46, η^2_p = .01], nor was the main effect of grade [$F(1, 96) = 3.96$, p = .05, η^2_p = .04]. As there were no significant differences between grade cohorts, the ANOVA on $RAVLT_{t1-5}$ scores was collapsed across grade. There was no significant effect of condition on $RAVLT_{t1-5}$ scores [$F(1, 96) = 1.79$, p = .18, η^2_p = .02).

With regard to PSI scores, the 2 (instruction: withhold or best effort) x 2 (grade) ANOVA showed that the interaction between grade and instructions was not significant

($F(1, 96) = 1.11$, $p = .30$, $\eta^2_p = .01$], nor was the main effect of grade $F(1, 96) = 1.96$, $p = .17$, $\eta^2_p = .02$]. When collapsing across grade, the main effect of instructions was not significant ($F(1, 96) = 2.12$, $p = .15$, $\eta^2_p = .02$]. It is important to note that although there was no significant effect of instruction given, it is possible that some participants within the withholding group did withhold, however, their performance may have been masked by those who did not withhold.

Question 2: Can the TOMM and MSVT be used to detect children who are withholding their best effort?

The mean and standard deviations for scores on the MSVT are presented in Table 25. It is particularly interesting to note that despite the lack of between group differences on the PSI and $RAVLT_{t1-5}$ based on instructions given, there remains large between group differences in the variances of the MSVT scales. Collapsed across grade, all p's < .01 except for the FR scale.

The mean and standard deviations for scores on the TOMM are presented in Table 26. Since participants in Study 1 did not complete the TOMM, there is no comparison group to which I can compare the results of participants who were instructed to withhold. Thus, the only psychometrics I can calculate regarding classification accuracy of the TOMM is the sensitivity.

Table 25. Mean (SD) MSVT scores across grade and instruction given for participants in the withholding group and comparison group

			IR	DR	CO	PA	FR
Grade 4	Comparison Group	*n* = 20	99.25 (2.45)	99.75 (1.11)	98.50 (3.29)	99.50 (2.24)	72.25 (10.06)
	Withholding Group	*n* = 20	96.50 (11.48)	96.00 (11.77)	94.50 (15.04)	97.00 (11.29)	71.00 (16.03)
Grade 6	Comparison Group	*n* = 32	99.22 (2.24)	99.38 (1.68)	98.91 (2.45)	97.81 (5.53)	73.44 (15.58)
	Withholding Group	*n* = 28	97.32 (7.99)	95.71 (11.84)	94.82 (14.37)	96.07 (9.17)	76.43 (14.58)

Table 26. Mean (SD) TOMM Scores for Participants in the Withholding Group as a Function of Grade

		Trial 1	Trial 2
Grade 4	*n* = 20	45.45 (4.99)	48.45 (4.70)
Grade 6	*n* = 28	45.93 (4.08)	49.11 (2.60)

Table 27. Correlations between subscales of the MSVT among participants in the withholding condition as a function of grade

		IR	DR	CO	PA	FR
Grade 4	IR	1				
n = 20	DR	.98**	1			
	CO	.97**	.99*	1		
	PA	.95**	.90**	.84**	1	
	FR	.33	.30	.35	.34**	1
Grade 6	IR	1				
n = 28	DR	.95**	1			
	CO	.96**	.99**	1		
	PA	.76*	.83**	.82**	1	
	FR	.19	.30	.27	.40	1

* Significant at the $p < .01$ level

** Significant at the $p < .001$ level

As the MSVT subscales were strongly related (see Table 27), the appropriate analysis would be a MANOVA, however, the assumptions for a MANOVA are not met: For four of the five MSVT subscales, Levene's test of homogeneity of variances showed unequal variances (excluding FR scale, all p's < .001). Scores on all five subscales for both grade cohorts were negatively skewed to varying extents and in some cases exceeded acceptable cut-offs (skewness ranged from -0.27 to -3.73). In addition to the

scales being skewed, they were not normally distributed (Shapiro-Wilk *p*'s ≤ .001 for the IR, DR, CO, and PA scales, and *p* = .02 for the FR scale). Visually, the distributions were unimodal and the outliers were not sufficiently distant (i.e. > 3 standard deviations from the mean) that a few data points are distorting the data. Given that the assumptions for MANOVA were not met and the sample sizes were not sufficiently large to buffer against violations of the assumptions, the appropriate non-parametric test is the Kruskal-Wallis H test. The results showed no significant difference between the grade cohorts for any of the MSVT subscales [$\chi^2(1)$ ranged from 0.01 to 1.79 with all p's > .05]. Scores between the grade cohorts were thus collapsed for the analysis on the effect of instruction. Across all five MSVT subscales, there were no significant between group differences based on instructions given [$\chi^2(1)$ ranged from 0.17 to 3.74, with all *p* values greater than .05 (see Table 28 for specific results).

Although there were no between group differences based on instructions given, the difference in variances between the groups suggests that there was indeed an effect, albeit not a statistically significant one. Among the participants in the comparison group, the standard deviations are in the range of 1.11 to 5.53 for the IR, DR, CO, and PA scales. Among the withhold group, the standard deviations are in the range of 7.99 to 15.04. On most scales, there is a 5 times increase in the magnitude of the variances across both grade cohorts (see Table 25). Indeed, the homogeneity of variance test reported above as part of assumption checking confirms that the variances are not statistically the same.

Table 28. Comparison of MSVT subscale scores between instruction groups

MSVT Scale	Chi Square	Cohen's *d*
IR	$\chi^2(1) = 0.35$, $p = .56$	0.33
DR	$\chi^2(1) = 3.74$, $p = .05$	0.44
CO	$\chi^2(1) = 1.06$, $p = .30$	0.39
PA	$\chi^2(1) = 0.34$, $p = .56$	0.26
FR	$\chi^2(1) = 0.17$, $p = .68$	-0.08

n = 100

To assess the classification accuracy of the MSVT and TOMM based on assigned instruction group, I first categorized each participant as having passed or failed each test according to the test publisher's user manual. For the proportions of participants that were classified as true positives, true negatives, false positives, and false negatives, see Table 29 for the MSVT. For the proportions of true positives and false negatives on the TOMM, see Table 30. Note that the proportions of participants in each cell were almost identical between the grade cohorts for both the MSVT and TOMM and they were thus collapsed by grade.

Table 299. Proportion of participants that were classified as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) by the MSVT

	Withhold Group	**Comparison Group**
Fail MSVT	TP = 10.4% (*n* = 5)	FP = 0% (*n* = 0)
Pass MSVT	FN = 89.6% (*n* = 43)	TN = 100% (*n* = 52)

Table 30. Proportion of participants in the withholding group that were classified as true positives (TP) and false negatives (FN) by the TOMM

Fail TOMM	TP = 4.3% (*n* = 2)
Pass TOMM	FN = 95.8% (*n* = 46)

As can be seen in tables 29 and 30 above, the base rates for some cells were particularly small. As a result, it would not be accurate to calculate the psychometric properties of the test as they would be unreliable. By this, I mean that a difference of one participant may change a metric substantially, such as the sensitivity from 1.00 to 0.50. Thus, one would have little confidence in the precision of these scores. Based on the data, however, one can observe that the vast majority of participants in both the withholding and comparison groups passed the MSVT and TOMM. As previously discussed, in the comparison group, there were no false positives. In the withholding

group, only 10% of the sample was detected as putting forth less than optimal performance. This may be explained by participants evading detection or not withholding their best effort as instructed.

As before, I was interested in the classification accuracy of the MSVT and TOMM according to whether participants actually withheld their best effort and not whether they were instructed to withhold their best effort. As above, a limitation of this approach is the lack of known group membership against which classification decision can be compared. Given the low base rates of participants who were categorized as "likely withholding", these statistics cannot be calculated as the results would be unreliable. In reviewing the ratios reported in table 31, it is important to note that both the RAVLT and PSI total scores for participants in the withholding condition of the current study are normally distributed [$RAVLT_{t1\text{-}5}$: Shapiro-Wilk = .95, p = .05, PSI: Shapiro-Wilk = .98, p = .70), and thus I would expect a base rate of approximately 16% of the participants to obtain scores at or below one standard deviation below the mean on each test. For the proportions of participants who were classified by the MSVT as true positives, true negatives, false positives, and false negatives, see Table 32.

Table 31. Proportions of participants in each group who obtained scores lower than one standard deviation below the mean on either the PSI or $RAVLT_{t1\text{-}5}$ as a function of grade and instruction given

		Comparison Group	Withholding Group
Grade 4	$RAVLT_{t1\text{-}5}$	20.0% (*n* = 4/20)	15.0% (*n* = 3/20)
	PSI	10.0% (*n* = 2/20)	45.0% (*n* = 9/20)
Grade 6	$RAVLT_{t1\text{-}5}$	9.4% (*n* = 3/32)	17.9% (*n* = 5/28)
	PSI	18.8% (*n* = 6/32)	17.9% (*n* = 5/28)

Table 322. Proportion of participants that were classified by the MSVT as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as a function of obtaining scores on either or both measures above or below the -1 SD cut-off

	PSI +/ $RAVLT_{t1-5}$ Score	**MSVT Fail**	**MSVT Pass**
Grade 4	≤ -1 *SD* cut-off	TP = 100.0% (*n* = 2/2)	FN = 36.8% (*n* = 14/38)
	> -1 *SD*	FP = 0.0% (*n* = 0/2)	TN = 63.2% (*n* = 24/38)
Grade 6	≤ -1 *SD* cut-off	TP = 66.7% (*n* = 2/3)	FN = 35.6% (*n* = 16/45)
	> -1 *SD*	FP = 33.3% (*n* = 1/3)	TN = 91.1% (*n* = 41/45

With regard to the TOMM, the sample only includes participants who were instructed to withhold. The proportions of participants that were correctly and incorrectly classified according to the cut-off of -1 *SD* on either or both measures are presented in Table 33.

Table 333. Proportion of participants that were classified by the TOMM as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as a function of obtaining scores on either or both measures above or below the -1 SD cut-off

	PSI +/ $RAVLT_{t1-5}$ Score	**TOMM Fail**	**TOMM Pass**
Grade 4	≤ -1 *SD* cut-off	TP = 100.0% (*n* = 1/1)	FN = 47.4% (*n* = 9/19)
	> -1 *SD*	FP = 0.0% (*n* = 0/0)	TN = 52.6% (*n* = 10/19)
Grade 6	≤ -1 *SD* cut-off	TP = 0.0% (*n* = 0/0)	FN = 37.0% (*n* = 10/27)
	> -1 *SD*	FP = 100.0%	TN = 63.0%

	(*n* = 1/1)	(*n* = 17/27)

As previously discussed, the classification properties for the TOMM cannot be reliably computed with this data set due to low base rates in some cells.

Additional analyses. As in Study 2, I was interested in whether failure on at least one PVT was associated with lower scores on the ability-based measures compared to those who passed both PVTs. This is important as the premise of using PVTs is to infer how the examinee approached the testing session, notwithstanding that optimal performance can be variable throughout the testing session. To examine whether there were between-group differences based on having failed at least one PVT, I classified participants as having passed both PVTs or having failed at least one of them. This resulted in 43 participants in the group that passed both PVTs, and 5 participants in the group that failed at least one PVT. The unequal sample sizes posed a problem for meeting the assumptions to conduct an ANOVA and there is insufficient power to detect an effect. However, the mean group scores are illustrative of the between-group differences and are reported in Table 34 below.

Table 344. Mean (SD) PSI and $RAVLT_{t1-5}$ scores for participants who passed both PVTs or were detected by at least one PVT

	PSI	$RAVLT_{t1-5}$
Passed Both Effort Tests (*n* = 43)	102.02 (18.11)	41.98 (15.85)
Failed at Least One Effort Test (*n* = 5)	86.40 (21.30)	35.23 (17.65)

The data suggests a medium/large mean difference on the PSI (cohen's *d* = 0.79) and a small/medium mean difference on the $RAVLT_{t1-5}$ (cohen's *d* = 0.40) based on whether the participants passed both PVTs or failed at least one of them, however, this should be interpreted cautiously given the small and unequal cell sizes. While the group of participants who failed at least one PVT is small, it remains worthwhile to examine whether the PVTs can be used to detect participants who put forth suboptimal effort.

I further examined the characteristics of the group of participants instructed to withhold who passed the PVTs. Among the fourth grade participants instructed to withhold, 18/20 passed both PVTs. Of this group who passed the PVTs, eight

participants obtained a score on at least one measure that was at least one standard deviation below the comparison group mean. Of those eight participants, only one obtained a score below the same threshold on both the $RAVLT_{t1-5}$ and PSI. Of the scores below the threshold, only one $RAVLT_{t1-5}$ and one PSI score exceeded two standard deviations below the comparison group mean, thus the majority of the scores below the threshold were between one and two standard deviations below the group mean.

Among the sixth grade participants, 25/28 passed both PVTs. Of the group who passed the PVTs, 8 participants obtained scores on at least one measure that was lower than one standard deviation below the comparison group mean. None of the participants obtained a score below this threshold on both the $RAVLT_{t1-5}$ and PSI. Only one participant obtained a score that exceeded two standard deviations below the mean and thus almost all of the scores below the one standard deviations threshold were between one and two standard deviations below the comparison group mean.

To examine whether there were differences between children in grade four and grade six within these subsamples of participants who were instructed to withhold but passed the PVTs, I used two separate one-way ANOVAs with grade as the between-subjects factor and scores on the PSI and RAVLT as the DVs. There were no differences between grades on either the PSI [$F(1, 41) = 1.50$, $p = .23$] or RAVLT [$F(1, 41) = 0.43$, $p = .52$].

We can once again compare the proportion of participants who passed both PVTs and obtained scores lower than the one standard deviation below the group mean on either the $RAVLT_{t1-5}$ or PSI to the proportion of participants in the comparison group who met the same criteria to see whether the distributions differed (H_0: $p_1 = p_2$). In the withholding condition 16/43 or 37% of the sample obtained a score below the threshold. In the comparison group, 14/52 or 27% of the participants met the same criteria. In testing whether the proportions are the same, I obtained a $z = 1.07$, $p = .28$, and thus failed to reject the null.

As in Studies 1 and 2, I examined whether participants' memory for the instructions was related to their performance. In the present study, I once again correlated number of instructions recalled and scores on the $RAVLT_{t1-5}$ and PSI for participants in the withholding group. I collapsed across grade as there is no indication of between grade differences, and increasing the number of correlations would increase my error rate and decrease my ability to detect an effect given the very few number of

participants who appeared to withhold their best effort. The relationship between memory for the instructions and scores on the $RAVLT_{t1-5}$ was not significant (r = .08, p = .61), nor was it significant for PSI scores (r = -.03, p = .84).

4.3. Discussion

In both the first and second studies, children's memory of the instructions was confirmed prior to the testing session to ensure participants knew how to respond to the test items. This was an important step in these studies to rule out the possibility that the participant did not understand or remember the instructions, however, it also reduced the ecological validity of the studies. To address this issue, I conducted a third study that was almost identical to the second study but with the exception of not confirming the instructions given to the children before the testing session. As the examiner did not confirm the instructions, participants were also unaware that the examiner knew they had been instructed to withhold their best effort. Other than this procedural change, this study was conducted identically to Study 2. As before, due to time constraints with the end of the school year, participants in Study 3 were only instructed to withhold their best effort and their scores are being compared to participants from the first session of Study 1. Thus, as there was no random assignment, these results should be interpreted with caution.

The first hypothesis was that children instructed to withhold their best effort would obtain lower scores on both the $RAVLT_{t1-5}$ and PSI than children who were instructed to try their best. There were no significant differences between grade cohorts, so scores were collapsed by grade. Comparing the group instructed to withhold their best effort to the group instructed to try their best, there were no significant differences between them on either the $RAVLT_{t1-5}$ or PSI.

The second hypothesis was that the TOMM and MSVT could be used to detect children who were withholding their best effort. However, as there were no significant between-group differences and thus the withholding group as a whole did not withhold their best effort, I expected participants to pass both PVTs. The results showed that 10% of the fourth grade participants and 11% of the sixth grade participants failed the MSVT. This is consistent with the larger variances on the MSVT among those instructed to

withhold compared to those who were instructed to try their best. On the TOMM, 5% of the fourth grade participants and 4% of the sixth grade participants were classified as performing suboptimally. While these rates are relatively small, all of the children instructed to try their best passed the MSVT and other studies have shown 100% pass rates on the TOMM for children instructed to try their best (Constantinou & McCaffrey, 2003; Donders, 2005; Gast & Hart, 2010; Kirk et al., 2011; MacAllister et al., 2009; Rienstra et al., 2010; Schneider, Kirk, & Mahone, 2014). Thus, it appears that only a small minority of participants followed the instructions to withhold their best effort and were detected by the PVTs.

With regard to the classification accuracy of the MSVT and TOMM when instructions given was the criterion, there was an insufficient number of participants in some cells to calculate the test statistics due to there being too few participants who were detected by the PVTs as performing suboptimally. Qualitatively, the test appeared to perform well at correctly passing participants who put forth their best effort. It is unclear, however, whether the PVTs can correctly detect those who were performing suboptimally as many participants did not follow the instructions given to withhold. The alternative explanation that participants evaded detection is not supported by the data as the groups performed statistically equally on the PSI, RAVLT, and PVTs. It is noteworthy that the results were nearly identical between grade cohorts. It remains possible that there is a more efficient cut-off score to use with children than the cut-off score provided by the test publisher for adults and used in other studies, but there is insufficient data in this study to calculate such a cut-off and it is beyond the scope of this dissertation.

As before, I was interested in the classification accuracy of the PVTs among those who actually withheld rather than those who were instructed to withhold. However, since few, if any, participants actually withheld their best effort, I was unable to calculate accurate and reliable test statistics.

As in Study 2, I was interested in whether participants who failed at least one PVT obtained lower scores on the $RAVLT_{t1\text{-}5}$ and PSI than those who passed both PVTs. Using this criterion, I divided the participants who received the withholding instructions into two groups: passed both PVTs or failed at least one. Unfortunately, this yielded very small sample sizes for the groups of participants who failed at least one PVT and I could not conduct any inferential statistics. However, looking at the mean scores, there was a

clear trend for participants who were detected as putting forth suboptimal effort to achieve lower scores on the $RAVLT_{t1-5}$ and PSI than those who passed both PVTs.

I conducted a follow-up analysis of participants who were instructed to withhold their best effort but were not detected by either PVT. Among the fourth grade participants instructed to withhold, 18/20 passed both PVTs. Among the sixth grade participants, 25/28 passed both PVTs. Of the participants that were classified as likely withholding, the majority obtained scores between 1 and 2 standard deviations below the comparison mean on either or both the RAVLT and PSI. Among this sample of participants who were instructed to withhold their best effort but passed the TOMM and MSVT, there were no significant differences between the grade cohorts on either the $RAVLT_{t1-5}$ or PSI. As a group collapsed by grade, I found that the proportions of participants that obtained scores lower than one standard deviation below the comparison group mean was not statistically different than the proportion of participants that obtained similar scores in the comparison group. Thus, the data suggests that these participants as a group were unlikely to be withholding and had tried their best or withheld so minimally that it was not detected.

Lastly, I wanted to examine whether memory for the instructions after the testing session was related to performance. The relationship between memory for the instructions and performance was not significant for the $RAVLT_{t1-5}$ or PSI, however, this is not surprising given the non-significant between-group differences and the apparently small number of participants who may have actually been withholding their best effort.

Chapter 5. General Discussion

The sequence of studies in this dissertation progressed from examining whether children had the capacity to withhold their best effort under optimal and controlled conditions to a design where the testing session more closely resembled a clinical testing session. In the first study, the examiner ensured that participants remembered and understood the instructions before beginning the tasks. In this study, I used a repeated measures design to reduce error due to individual variability and calculated change scores for a measure of the degree of withholding. While this allowed for a sensitive measure of withholding, the study lacked external validity as the majority of psychological tests that are used in this area of practice are protected from the public domain and children would not have had prior exposure to the testing materials before being asked to withhold. Secondly, the generalizability of the results may be questioned as another study (Nagle et al., 2006) has shown that children have the ability to withhold their best effort, but only after trying their best on their first exposure to the tests. This begged the question of whether the results from this study would generalize to the real world. I took the next step of modifying the research design to only include a single session to see whether children could still withhold their best effort, even without prior exposure to the tests. While this addressed the question of needing prior exposure to the testing materials before withholding, the procedure was still artificial as the examiner reviewed the instructions for participants to withhold their best effort prior to beginning the tasks and ensured that participants understood how they were supposed to perform on the tests. Thus, for the third study, I modified the procedure once more by removing some of the checks that let me control the influence of extraneous variables in the first two studies, thus making the testing session more like a “real world” testing session. In this third study, the examiner did not remind the child of the instructions nor ensure that the child understood the instructions prior to beginning the tasks. In this last study, participants did not know whether the examiner was aware that they had been instructed to withhold their best effort.

Before I compare the results from the three studies, I want to state the limitations of making these cross-study comparisons (other limitations of the studies are discussed below). In the first study, participants were randomly assigned to an experimental (withholding) or control (optimal effort) condition. Due to time constraints, in the second and third studies, participants were only instructed to withhold their best effort and the results were compared to participants who had put forth their best effort in Study 1. The potential problem is that since participants were not randomly assigned, it is possible that the group of participants in Studies 2 and 3 were systematically different from the comparison group in Study 1. If the samples were not equivalent, then differences between the groups may not be attributable to the experimental manipulation. There are at least two systematic differences between the groups that need to be considered. First, Study 1 took place earlier in the school year than Studies 2 and 3. While Studies 2 and 3 took place 2-3 months after Study 1, we know that children's cognitive abilities develop rapidly and for some tests, normative scores are divided into quarters of the year for that reason. Despite the slight difference in testing time during the year, the average age between the groups was not significantly different and any difference that may have existed due to age would have been small. The second factor to consider is that participants in Studies 2 and 3 were sampled from different schools than participants in Study 1. While the schools were different, participants were all from the same school district and all of the schools were in the greater Vancouver area. While this is certainly not ideal, it is important to note that the tasks in these studies assessed cognitive abilities, which generally develop with age and none of the tasks assessed specific knowledge, which is more closely tied to learning in school. Nonetheless, it is possible that the students' cognitive abilities varied by school and some systematic difference may have confounded the results. Thus any comparisons between studies should be viewed with caution.

5.1. Children's Ability to Withhold

The first hypothesis for all three studies was that children who were instructed to withhold their best effort would perform worse on measures of memory and processing speed than children who were instructed to try their best. The first study was designed to

make it relatively easy for participants to follow the instructions to either withhold or put forth their best effort. In this study, there were large between-group differences on both the $RAVLT_{t1-5}$ and PSI based on the instructions given to the participants. The effect was present for both grade cohorts. In the second study, the task was slightly more difficult as participants did not have prior exposure to the testing materials. In this study, the only between-group differences based on instructions provided was among fourth graders on the PSI. It is noteworthy that in Study 1, there was a larger effect size on the PSI than the $RAVLT_{t1-5}$ between instruction groups. Thus, it is possible that it may have been easier for children to perform the task more slowly than to withhold responses. Aside from the difference on the PSI among fourth graders, there were no between-grade cohort differences in either of the first two studies.

In the second study, in addition to the effect of instruction on PSI scores among fourth graders, there was also an interesting effect of instructions on MSVT scores. Between the two groups, there was no statistically significant main effect of instruction on MSVT scores overall, however, the variances between the groups were dramatically different and are statistically different (as per Levene's test). Among both groups, the mean scores were at the ceiling of the range for the IR, DR, CO and PA subscales, however, among the comparison group, there was a narrow range of variance around the mean and among the withholding group, the variance was quite wide; on some subscales, the variances were ten times larger in the withholding group than the comparison group. On review of the raw data, the large variances among those instructed to withhold was not attributable to a small number of outliers, but rather the scores were negatively skewed and continuous. Thus, there was clearly an impact of the instructions given on how some participants approached the testing session, even though the main effect of instruction on PSI and $RAVLT_{t1-5}$ scores was not significant. There are a few possible explanations for this: (1) Participants only withheld on particular tests. If this were the case, it would likely have been random because participants ought to have tried their best on the PVT and performed poorly on the PSI and $RAVLT_{t1-5}$, rather than vice versa. (2) Participants withheld their best effort so minimally that it was not detected by the primary ANOVA used to test for between group effects based on instructions. This is possible, but one would expect at least a small difference between the withholding and comparison group means, or alternatively a negative skewing of scores based on the degree of withholding exhibited. (3) Another possible explanation is

that the MSVT is more sensitive to withholding than the $RAVLT_{t1-5}$ or PSI, and thus detected subtle withholding whereas the PSI and $RAVLT_{t1-5}$ did not. (4) Lastly, that the MSVT as a test is less cognitively demanding than the other tests and freed up enough cognitive resources for some participants to implement the process of identifying the correct answer and then choosing a deliberatively incorrect response. This would also explain why some children were better able to withhold their best effort, or could withhold to a greater degree on the PSI rather than the $RAVLT_{t1-5}$. Being a forced choice test, the MSVT may be easier to perform poorly on because the examinee does not have to generate an alternative response, the examinee must simply choose from the incorrect options available. Similarly, it may be easier to withhold one's best effort on the PSI because the examinee must only perform slower, rather than go through the process of identifying the correct response and then generating an alternative incorrect answer. The importance of cognitive load and the differential ability of examinees to withhold based on the nature of the task could be further explored in a future study, as discussed below.

In the third study, the experimental testing session most closely resembled a real, clinical testing session and participants were not given any reminders about the instructions they needed to follow. In this study, there were no differences on either the PSI or $RAVLT_{t1-5}$ between the groups based on the instructions provided. However, the same pattern of results was observed on the MSVT, although somewhat attenuated compared to Study 2. In this study, the magnitude of the variance among the withholding group was approximately five times larger than the magnitude of the comparison group. The difference between the variances remained statistically different.

Collectively, the results show that some children have the capacity to withhold their best effort on some tests and under certain conditions. When the level of coaching was high (i.e. parents instructed their children and children were reminded of the instructions prior to the testing session), and children had prior exposure to the testing materials, children could withhold their best effort on the $RAVLT_{t1-5}$, PSI, and MSVT. Without prior exposure to the testing materials, however, only fourth grade children could withhold on the PSI and some children withheld on the MSVT. This was clearly a more demanding task and some children were unable to follow the instructions to withhold their best effort. In the last study, when the testing session most closely resembled a clinical testing session, children could not withhold their best effort on either the RAVLT or PSI, but the magnitude of the variances on subscales of the MSVT were different

based on the instructions given. Overall, the general trend was that as the task became more difficult, the impact of the instructions on performance dwindled, but there remained a small effect of the instructions given on how children approached the testing session and performed, most notably on the MSVT.

The pattern of results from these three studies is consistent with other simulation studies that have shown that children have the ability to withhold when then the task is simple, but their success at following the instructions to withhold wavers as the instructions become more complicated and the task becomes increasingly cognitively demanding. For example, when children are given simplified instructions to withhold, such as telling children to get as many answers wrong as possible (e.g. Gunn, Batchelor, & Jones, 2010), there are high failure rates on PVTs. When examiners use more complex instructions, such as asking children to get some questions wrong, but not too many (e.g., Blaskewitz et al., 2008), failure rates on PVTs are lower. In the present studies, the biggest differences between the groups based on instructions given was in Study 1 when the task was easiest and the differences between groups diminished progressively in Studies 2 and 3 as the task became increasing difficulty.

The question arises, what differences between the three studies might explain the observed differences in children's ability to withhold on the ability-based measures? Previous simulation studies have shown that prior exposure to the testing materials was not necessary for participants to withhold (e.g., Blaskewitz et al., 2008; DeRight & Carone, 2015; Gunn, Batchelor, & Jones, 2010). However, as the effect sizes between the groups based on instructions given were larger in Study 1 than in Studies 2 and 3, prior exposure to the tests may have made it easier for children to withhold. This is consistent with the findings by Nagle and colleagues (2006) who found that prior exposure to the testing materials facilitated children's successful withholding of their best effort. The reason why it may have been easier for children to withhold when they had seen the tests, as in Study 1, is that they were better able to develop a strategy having already seen the test or were familiar with the format of the testing session which freed up cognitive resources that could be devoted to remembering the instructions to withhold and to implement their withholding strategy. It is also possible that for some children, there was a strong need to demonstrate their best effort before inhibiting their responses as children are taught to put forth their best effort and are typically keen to show their skills. This is especially true in a school setting, which is where the study took place.

One of the key differences between the first two studies and Study 3 was whether the examiner reminded the child of the instructions prior to the testing session. While I assumed that parents were truthful in confirming that they had coached their child to withhold, if they did not, then children only received the instructions from the experimenter in Studies 1 and 2, and not at all in Study 3. In Study 1, memory for the instructions was significantly related to scores on the PSI and $RAVLT_{t1-5}$, and in Study 2 memory for the instructions was significantly related to PSI scores (which was the only significantly main effect of instruction group). In Study 3, scores on the $RAVLT_{t1-5}$ and PSI were not significantly related to memory for the instructions, albeit there was no significant effect of instructions on PSI or $RAVLT_{t1-5}$ scores upon which to correlate number of instructions remembered. The data does confirm that at least some children in Study 3 could recall the instructions after the testing session, but there were far fewer who could recall the instructions than in Studies 1 and 2. The low number of instructions recalled and attenuated range of number of instructions recalled could explain the absence of a significant relationship between number of instructions remembered and performance on the PSI and $RAVLT_{t1-5}$. While not entirely consistent between all three studies, the data suggests that memory for the instructions was a relevant factor in children's ability to withhold.

Overall, the data shows that children have the ability to withhold their best effort, but they require sufficient instruction and reminders prior to the testing session. The data also indicates that prior exposure to the testing materials generated more consistent and robust between-group differences on the ability-based measures. While the tests are protected from the public domain, it is not inconceivable that parents could obtain a copy of the tests or at least sample test items that would help children decrease the cognitive load during the testing session and/or generate a strategy for how to perform on testing. Thus, the evidence supports that some children can withhold their best effort if certain conditions are met.

5.2. Relationship between Withholding and Inhibitory Control

The second hypothesis from Study 1 was that individual differences in children's ability to withhold would be related to their scores on measures of inhibitory control. To measure inhibitory control, I used the Stroop and the Inhibition subtest from the NEPSY, the latter being similar to the Stroop but non-verbal. Neither measure of inhibitory control, individually or collectively, explained a significant amount of variability in change scores among the group of participants instructed to withhold. There are a number of possible explanations for the null finding. First, although it makes theoretical sense that inhibitory control would be related to withholding, it is possible that these variables are unrelated. It is also possible that the relationship is subtle and I did not have sufficient power to detect the effect.

The assumption in finding a relationship between inhibitory control and withholding is that there would be sufficient variability in both constructs to detect the relationship. It is possible that among participants, inhibitory control was sufficiently developed that there was not enough variability to show the relationship. However, a review of the standard deviations of the measures of inhibitory control and performance on the PSI/ $RAVLT_{t1-5}$ (see Table 1) shows that the standard deviations for scaled scores and T-scores were generally consistent with what would be expected in the general population (i.e., scaled score *SD* ~ 3 and t-score *SD* ~ 10) and in some cases exceeded what would be expected.

Earlier in this dissertation, I discussed a common issue with measuring executive functions known as the task impurity problem. In brief, it is inherently difficult to measure executive functions in isolation as the task of measuring a specific executive function will invariably draw on other executive functions or skills. For example, in the Stroop task, in addition to inhibitory control, one is also measuring reading, processing speed, verbal expression, visual scanning, attention, task switching, and working memory. Thus, measures of inhibitory control are impure and one is actually measuring multiple constructs. While both measures of inhibitory control attempt to isolate the construct of inhibition by calculating an interference score, the process of measuring inhibition will always capture other executive functions. This could account for a null finding because it

is difficult to isolate the variable of interest and it is not clear to what extent the null finding is attributable to the variance that is not associated with inhibitory control. A future study may consider the method proposed by Friedman et al. (2008) to reduce the impact of task impurity on measuring a specific executive function by statistically extracting the variance shared by multiple tasks that have different and/or non-executive requirements but share the same underlying latent ability (i.e., inhibitory control). This method reduces the degree of measurement error to isolate the variable of interest. Related to this, it may also be interesting to examine whether some serial combination of executive functions explains individual differences in withholding, such as the combination of inhibitory control, set-shifting, and working memory.

Anecdotally, it appeared that some children withheld their best effort for too many responses while others did not withhold their best effort at all. This pattern may also be explained by difficulty in task switching as participants would be required to switch between getting the answer correct and getting the answer incorrect. A future study may explore the role of task-switching on children's ability to withhold.

As I showed in my post-testing session questions, memory for the instructions at the end of the study was a significant predictor of withholding on both the $RAVLT_{t1-5}$ and PSI in Study 1, and on the PSI in Study 2. It is unclear whether this implies that children's memory of the instructions themselves was predictive, or whether the children's memory of the instructions simply reflects their short-term memory. Logically, it would make sense that remembering the instructions themselves would be essential and that greater memory of the instructions may be related to better developed short-term memory. In the future, it would be interesting to measure short-term memory as a separate ability unrelated to the task to see whether this ability plays an important role in withholding. In addition to considering the role of short-term memory, it would also be interesting to examine the role of working memory in children's ability to withhold. Certainly, it makes intuitive sense that holding an instruction set in mind while thinking of the correct answer and then inhibiting that response in place of an alternative response would rely on working memory. Moreover, the task of self-monitoring the number of correct and incorrect responses as to not withhold too much or too little would also rely on working memory and may implicate this construct in withholding.

Study 1 was the first study to examine the question of why some children may be able to withhold their best effort while others cannot. Gombos (2006) hypothesized that

inhibitory control is an important cognitive ability in deception and that increased cognitive load may impair inhibitory control. Certainly, withholding one's best effort is more difficult than trying one's best and limited cognitive resources may explain why some children were unsuccessful at withholding. The difference between those who could withhold and those who could not may be better explained from the perspective of limited cognitive resources given the complicated demands of the task. It is certainly possible that changes in the complexity of the task and the associated cognitive load may explain the differences in children's ability to withhold across the current studies.

5.3. The Effectiveness of PVTs to Detect Withholding

To assess whether the PVTs performed well at discriminating between children who withheld from those who tried their best, I used two different criteria to categorize children as withholding. The first was by the instructions given to their group and the second was by whether the participants were likely withholding based on their performance on the $RAVLT_{t1-5}$ and PSI. I included this second categorization as I was interested in whether the PVTs detected children who actually withheld, not just those who were instructed to withhold.

Using the instructions given to participants as the criterion, the first study showed large between-group differences on all five MSVT scales. None of the participants instructed to perform their best were classified by the MSVT as putting forth suboptimal effort, but approximately half of the participants instructed to withhold were not classified as such. In the second study, the MSVT accurately classified approximately the same proportion of participants as Study 1 who were instructed to withhold their best effort. In this study, I also administered the TOMM and it classified even fewer participants who were instructed to withhold as putting forth suboptimal effort. In the third study, where participants were not provided with a reminder of the instructions prior to the testing session, both the MSVT and TOMM accurately classified even fewer participants than Study 2 as having been instructed to withhold their best effort. This either suggests that the group of participants in this study were better at evading detection than participants in Studies 1 and 2 (which is unlikely as the instructions were the same in all three studies), or alternatively, as the tasks became more difficult, fewer participants followed

the instructions to withhold and in fact tried their best. If the latter explanation were true, then using the instructions given to participants to classify them as withholding or trying their best may have resulted in some participants being misclassified, as their group membership may be better defined by the way they performed on testing (i.e., withholding or trying their best) rather than by the instructions they were provided. In addition, it suggests that the metrics calculated to evaluate the accuracy of the PVTs to detect withholding cannot be relied upon.

For the second set of analyses, I categorized participants as likely withholding their best effort or not based on their performance on the $RAVLT_{t1-5}$ and PSI, however, this resulted in very low base rates within some cells, which did not allow for accurate calculation of PVT metrics. In epidemiological studies where there is a low base rate of something being detected, this problem is overcome with extremely large data samples, but in the current study, the samples were not nearly big enough to overcome this problem (Woodward, 2005). In general, the test performed comparably using the performance based criterion compared to assigned condition in the first study, but there was a greater number of false positives from the control condition. This is likely attributable to a small number of participants obtaining lower scores on the second testing session than the first as part of normal test-retest variation in scores. In the second and third studies, I faced the same issue regarding low base rates and fewer participants were detected as withholding, which can likely be explained by fewer participants actually withholding as noted elsewhere.

Overall, these results suggest that using instructions provided or a performance-based cut-off as the criterion to distinguish between withholders and those trying their best are both problematic ways to define group membership for the purpose of assessing the classification accuracy of PVTs with children. When using instructions provided as the criterion, one cannot assume that participants followed the instructions. When using a performance-based criterion, the PVTs may misclassify participants who were trying their best but performing below average as likely withholding.

As a supplemental analysis, I examined the group of participants who were instructed to withhold their best effort but passed the PVTs. The results showed that these participants obtained scores that were very similar to the comparison group of participants who tried their best and were thus unlikely to have been withholding their best effort. In other words, the participants who were instructed to withhold their best

effort and passed both PVTs performed comparably to those who were instructed to try their best, which suggests they did not evade detection, but rather did not follow the instructions to withhold and tried their best.

While failure on a PVT is not synonymous with withholding, fewer participants were detected as likely withholding their best effort as the task became more difficult. The instructions given to children in the withholding conditions were the same in all three studies and children in the second and third studies should not have been any better at evading detection than those in the first study. The overall difference between the studies was the level of difficulty in remembering and implementing the instructions to withhold. It would not make sense for children to get better at evading detection as the task became more difficult. Thus, the decline in the accuracy of the PVTs to detect withholding likely reflects fewer participants withholding their best effort rather than participants being better at evading detection.

The first study in this series showed that when children are instructed to try their best, they consistently pass PVTs with perfect or near perfect pass rates. Examining whether PVTs can detect children who are withholding is far more difficult as one cannot be certain that a child is following the instructions to withhold. Without knowing whether a participant is actually withholding their best effort, one cannot accurately evaluate whether the PVT can accurately detect them.

5.4. Relationship between Failing a PVT and Ability-Based Test Scores

The principle behind using PVTs in clinical practice is that an examinee's approach to testing on one test may reflect their overall approach to testing and should be considered in the interpretation of other tests. I examined whether failure on a PVT was associated with lower performance on the $RAVLT_{t1-5}$ and PSI. In Study 1, there was a clear relationship between failure on a PVT and performance on the ability-based measures. In Study 2, there were significant differences on the $RAVLT_{t1-5}$ and PSI for sixth grade students based on whether they passed or failed at least one PVT, but the same effect was not significant among fourth graders. For the third study, the grade cohorts were collapsed due to the small number of participants who failed at least one

PVT and I was unable to compare the groups statistically, however, the mean scores show a clear pattern of differences on the PSI between those who failed at least one PVT and those who passed both PVTs.

The relationship between failing a PVT and performance on ability-based tests in the current studies is consistent with the majority of pediatric and adult studies which show that failure on PVTs is associated with a decrease in scores across various cognitive domains (e.g., Kirkwood et al., 2012). There is one study which dissents from the others in the literature, which was conducted by Perna and Loughan (2014), who found that children who failed the TOMM did not score differently than those who passed the TOMM on the majority of neuropsychological tests administered. They argue that failure on a PVT is not necessarily predictive of poor performance on other ability-based measures. While this study stands alone in the literature on pediatric malingering, it is still consistent with recommendations to not rely on a single PVT. Administering a PVT provides a "snapshot" of performance validity at a point in time in the testing session, however, performance validity may be variable throughout the testing session. As such, it is recommended and common practice for examiners to administer more than one PVT at different points within their testing session (Bush et al., 2005; Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009).

5.5. Limitations and Future Directions

As with any simulation study, there are a number of important limitations to discuss. One of the key differences between a simulation study and the real world is the participants' level of motivation. Among adults undergoing clinical assessments where secondary gain is at stake, it is well known that a significant proportion of examinees will exaggerate or fake symptoms to obtain compensation. Logically, this would extend to parents wanting their children to exaggerate or fake symptoms for financial gain. One would expect that in a real world scenario, parents would expend considerable time and effort researching how children should perform on testing in order to convincingly appear impaired. A real world testing scenario is dissimilar from a simulation study where there is very little at stake, such as the chance to win a cash prize for the parents or a toy for the children. In the present series of studies, parents read a brief book to their child the

night before testing and gave their child the instructions at the end of the book. Given that motivation was low and children were able to withhold under some circumstances, we can speculate that with a high level of motivation, parents and children would be better prepared to withhold their best effort and thus the results of these studies are possibly a low estimate of children's ability to withhold. Of course, it is also possible that with a high level of motivation, children may have performance anxiety and their performance may in fact decline. Future studies that examine withholding and factors that may impact performance may consider whether stress impacts children's ability to withhold.

It is worth mentioning that children in the third study were participating during the final two-three weeks of school before the summer break and may not have been fully engaged in the task; perhaps they would have preferred to be with their classmates as there are often fun activities planned for the end of the school year. Anecdotally, this did not appear to be the case and children did not ask to return to their classroom early (except for one), but it is nonetheless worth considering.

The assumption that parents followed the instructions to coach their child to withhold, as they confirmed on a signed form, is another potential limitation. It is possible that parents signed the form and did not review the storybook and instructions carefully with their child. If this were true, it may partially explain why so few children in Study 3 were identified as likely withholding. This may relate back to a parent's motivation to participate as there was far less incentive for them to coach their child than would be present in a real world scenario. I attempted to create an incentive by putting the names of parents who participated into a draw for a cash prize, but there was no way to verify that parents had indeed followed the instructions or had only signed the form to enter the draw.

Another limitation of the present studies was the location of where the testing took place. In all cases, children were removed from their classroom to participate in the testing session, but in some schools there were still distractions in the testing environment. For example, it was not uncommon to hear children walking in the hallways outside of the testing room. In a typical clinical setting, examiners ensure that the testing environment is free of distractions which is not possible in a school setting. In addition, given the short duration of the tasks, children may have been thinking of what was happening in their classroom rather than committing their attention to the task at hand.

An important consideration is that I was limited to conducting a brief testing session as teachers and parents did not want children removed from their classroom for an extended period of time. This was also done to reduce the chances of children requesting to return to their classroom before completing all of the tasks. In the real world, judgments about whether a child is withholding their best effort is based on patterns of results with consideration of various factors such as injury severity and pre-injury academic performance. In these studies, I had to make a determination of "likely withholding" based on very limited information. This is clearly a limitation as it would have been preferable to rely on more data to make that decision. It is also worth considering that withholding one's best effort in a clinical situation may be more difficult for children as they would have to remain consistent over a longer period of time or across multiple days of testing.

With regard for directions for future study, in the present studies, I controlled the way in which parents coached their children to withhold as I was interested in children's ability to withhold and did not want to allow the manner or degree of coaching to vary. To introduce more realism into a study, researchers could ask parents to gather their own information as to how they would coach their child to withhold. While this would be interesting, it does present some ethical and methodological issues that would have to be resolved.

As previously discussed, it would be interesting to examine the pattern of how much examinees withhold based on the task. In the present studies, there was a trend that the magnitude of withholding was greater on the PSI than the $RAVLT_{t1-5.}$ I hypothesized that this may be because it is far simpler to perform a task slower than to think of the correct response to a question and then select a different response. In addition, this may be related to and interact with the degree of cognitive load on the examinee. It is possible that examinees may find it more difficult to withhold when there is a greater cognitive load than when there is less demand on their cognitive resources.

Given what appeared to be varying degrees of withholding on the $RAVLT_{t1-5}$ and PSI, it would be interesting to use measures of effort that are embedded into other measures. This would help examine whether children detected which tests were measuring an ability and which were measuring effort. An example of this type of measure is the Reliable Digit Span (RDS) from the Wechsler scales. Chafetz (2008) has suggested combining data from effort tests with scores from other quantitative scales

(e.g., RDS) and qualitative aspects of test performance (e.g., Ganser like responses) on a rating scale called the Symptom Validity Scale (SVS). While there is little research to date, the data suggest that combining multiple parameters related to effort and malingering may be more useful than relying on a single score from effort testing alone. A future study may continue exploring the utility of combining various aspects of test performance to predict suboptimal effort. This may be especially interesting in children as their lies and deception may be less sophisticated than that of adults and may be easier to detect using a combination of quantitative and qualitative performance characteristics.

In terms of the results from the present studies informing future research, it is important to consider that measures of sensitivity, specificity, and predictive power may not reflect the psychometric properties of a test. In adult simulation studies on the detection of malingering, these metrics are used to describe the accuracy of the test and it is assumed that adults have the capacity to follow instructions to withhold their best effort. The results from the present studies, however, suggest that the psychometric properties of a test can vary based on the testing scenario and may also reflect children's ability to follow the instructions. Thus, using and relying on these types of test descriptors to assess the accuracy of a specific test to detect individuals who are withholding may be misleading as the apparent psychometric properties of the test may be highly contingent on the test procedure used in the study.

5.6. Implications for Clinical Practice

The results from these studies can be used to inform clinical practice by providing evidence that under certain conditions, some children have the ability to withhold their best effort, and PVTs may be useful at detecting withholding. Historically, it was assumed that children could not withhold, and when they did, it was easily detectable by the clinician. Data from the present studies and others shows that this previously held belief is outdated and children are better able to manipulate test results than previously thought. There is an emerging body of evidence that PVTs that have been used reliably and routinely with adults can also be used with children as young as 6 or 7 years old. The recommendation for testing adults remains the same as with

children, that clinicians use a multi-test and multi-method approach to assess performance validity. PVTs alone provides useful information to a clinician, but it must be considered within the broader context that includes interview data, mental status, behavioral observations, collateral information, the individual's history, neuropsychological / test performance, injury severity, potential for secondary gain, and any other available information (Walker, 2011).

As with adults, failure on a PVT should alert the examiner to take a closer look at the examinee's assessment results and scrutinize any inconsistent findings. It is encouraging that the false positive rate on the MSVT, TOMM, and other PVTs is virtually nil. Thus, there is a very low risk of mistakenly failing a PVT when trying ones best, however, participants who do not put forth their best effort may or may not be detected. Clinicians must also be mindful of the potential harm that may be caused by over testing in instances where a parent may be coaching a child to malinger and must guard against becoming complicit in a parent's effort to have a child unnecessarily assessed.

Based on these studies and others, I believe it is time for clinicians to begin considering the routine administration of PVTs in assessments of children, even when the possibility of secondary gain is remote, such as in assessments for learning disabilities, or where the child or their parent may receive some other benefit. While the results from PVTs may not be definitive, results can certainly assist the clinician by providing additional data to consider in formulating their opinion.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Anderson v. Wilson, ON CA 3753 (1999). Retrieved from http://canlii.ca/t/1f9kf

Anderson, V., Anderson, P. J., Jacobs, R., & Smith, M. S. (2008). Development and assessment of executive function: From preschool to adolescence. In V. Anderson, R. Jacobs, & P. J. Anderson (Eds.), *Executive functions and the frontal lobes: A lifespan perspective* (pp. 123-154). Philadelphia, PA: Taylor & Francis.

Anderson, V., Anderson, P. J., Northam, E., Jacobs, R., & Catroppa, C. (2001). Development of executive functions through late childhood and adolescence in an Australian sample. *Developmental Neuropsychology, 20,* 385-406. doi:10.1207/S15326942DN2001_5

Andrews v. Grand & Toy Alberta Ltd., 2 SCR 229 SCC (1978). Retrieved from http://canlii.ca/t/1mkb5

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87,* 84-94. doi:10.1207/s15327752jpa8701_07

Beardmore, S., Tate, R., & Liddle, B. (1999). Does information and feedback improve children's knowledge and awareness of deficits after traumatic brain injury. *Neuropsychological Rehabilitation, 9,* 45-62. doi:10.1080/713755588

Bigler, E. D. (1990). Neuropsychology and malingering: Comment on Faust, Hart, and Guilmette (1988). *Journal of Consulting and Clinical Psychology, 58,* 244-247. doi:10.1037/0022-006X.58.2.244

Birkich v. Cantafio, BCSC 40 (2016). Retrieved from http://canlii.ca/t/gmx5k

Blackwater v. Plint. SCC 58 (2005). Retrieved from Quicklaw.

Blandón-Gitlin, I., Pezdek, K., Lindsay, D. S., & Hagen, L. (2009). Criteria-based content analysis of true and suggested accounts of events. *Applied Cognitive Psychology, 23,* 901-917. doi:10.1002/acp.1504

Blandón-Gitlin, I., Pezdek, K., Rogers, M., & Brodie, L. (2005). Detecting deception in children: An experimental study of the effect of event familiarity on CBCA ratings. *Law And Human Behavior*, *29,* 187-197. doi:10.1007/s10979-005-2417-8

Blaskewitz, N., Merten, T., & Kathmann, N. (2008). Performance of children on symptom validity tests: TOMM, MSVT, and FIT. *Archives of Clinical Neuropsychology, 23,* 379-391. doi:10.1016/j.acn.2008.01.008

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10,* 214-234. doi:10.1207/s15327957pspr1003_2

Bond, C. F., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134,* 477-492. doi:10.1037/0033-2909.134.4.477

Bond, G. D. (2008). Deception detection expertise. *Law and Human Behavior, 32,* 339-351. doi:10.1007/s10979-007-9110-z

Boone, K. B. (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective.* New York, NY: Guilford Press.

Boone, K. B., & Lu, P. (2003). Noncredible cognitive performance in the context of severe brain injury. *Clinical Neuropsychologist, 17,* 244-254. doi:10.1076/clin.17.2.244.16497

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., … Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity: NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology, 20,* 419-426. doi:10.1016/j.acn.2005.02.002

Bussey, K. (1992). Lying and truthfulness: Children's definitions, standards, and evaluative reactions. *Child Development, 63,* 129-137. doi:10.2307/1130907

Carelli, M. G., Forman, H., & Mäntylä, T. (2008). Sense of time and executive functioning in children and adults. *Child Neuropsychology, 14,* 372-386. doi:10.1080/09297040701441411

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72,* 1032-1053. doi:10.1111/1467-8624.00333

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development, 69,* 672-691. doi:10.2307/1132197

Carone, D. A. (2008). Children with moderate/severe brain damage/dysfunction outperform adults with mild-to-no brain damage on the Medical Symptom Validity Test. *Brain Injury*, *22,* 960-971. doi:10.1080/02699050802491297

Carone, D. A. (2014). Young child with severe brain volume loss easily passes the Word Memory Test and Medical Symptom Validity Test: Implications for mild TBI. *The Clinical Neuropsychologist*, *28,* 146-162. doi:10.1080/13854046.2013.861019

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin, 113,* 403-439. doi:10.1037/0033-2909.113.3.403

Chafetz, M. D. (2008). Malingering on the social security disability consultative exam: Predictors and base rates. *The Clinical Neuropsychologist, 22,* 529-546. doi:10.1080/13854040701346104

Chafetz, M. D., Abrahams, J. P., & Kohlmaier, J. (2007). Malingering on the social security disability consultative exam: A new rating scale. *Archives Of Clinical Neuropsychology, 22,* 1-14. doi:10.1016/j.acn.2006.10.003

Chahal, K., & Cassidy, T. (1995). Deception and its detection in children: A study of adult accuracy. *Psychology, Crime & Law, 1,* 237-245. doi:10.1080/10683169508411959

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology, 23,* 201-216. doi:10.1016/j.acn.2007.08.010

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-scale deceit: Deception as a marker of two-, three-, and four-year-olds' early theories of mind. *Child Development, 60,* 1263-1277. doi:10.2307/1130919

Christoffersen v. Howarth, BCSC 144 (2013). Retrieved from http://canlii.ca/t/fvxw3

Clements v. Clements. SCC 32 (2012). Retrieved from QuickLaw.

Comalli, P. E., Jr., Wapner, S., & Werner, H. (1962). Interference effects of Stroop color-word test in childhood, adulthood, and aging. *Journal of Genetic Psychology: Research and Theory on Human Development, 100*, 47-53. doi:10.1080/00221325.1962.10533572

Constantinou, M., & McCaffrey, R. J. (2003). Using the TOMM for evaluating children's effort to perform optimally on neuropsychological measures*. Child Neuropsychology, 9,* 81-90. doi:10.1076/chin.9.2.81.14505

Conti, R. (2004). Malingered ADHD in adolescents diagnosed with conduct disorder: A brief note. *Psychological Reports, 94,* 987-988. doi:10.2466/PR0.94.3.987-988

Dash, J., & Dash, A. S. (1982). Cognitive developmental studies of the Stroop phenomena: Cross-sectional and longitudinal data. *Indian Psychologist, 1,* 24-33.

DeRight, J. and Carone, D. A. (2015). Assessment of effort in children: A systematic review. *Child Neuropsychology, 21,* 1-24. doi:10.1080/09297049.2013.864383

Donders, J. (2005). Performance on the Test of Memory Malingering in a mixed pediatric sample. *Child Neuropsychology, 11,* 221-227. doi:10.1080/09297040490917298

Edelstein, R. S., Luten, T. L., Ekman, P., & Goodman, G. S. (2006). Detecting lies in children and adults. *Law and Human Behavior, 30,* 1-10. doi:10.1007/s10979-006-9031-2

Ehrlich, S., Pfeiffer, E., Salbach, H., Lenz, K., & Lehmkuhl, U. (2008). Factitious disorder in children and adolescents: A retrospective study. *Psychosomatics: Journal of Consultation Liaison Psychiatry, 48,* 392-398. doi:10.1176/appi.psy.49.5.392

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist, 46,* 913-920. doi:10.1037/0003-066X.46.9.913

Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. *Psychological Science, 10,* 263-266. doi:10.1111/1467-9280.00147

Evans, A. D., Bender, J., & Lee, K. (2016). Can parents detect 8- to 16-year-olds' lies? Parental biases, confidence, and accuracy. *Journal of Experimental Child Psychology, 147,* 152-158. doi:10.1016/j.jecp.2016.02.011

Evans, A. D., & Lee, K. (2011). Verbal deception from late childhood to middle adolescence and its relation to executive functioning skills. *Developmental Psychology, 47,* 1108-1116. doi:10.1037/a0023425

Evans, A. D., & Lee, K. (2013). Emergence of lying in very young children. *Developmental Psychology, 49,* 1958-1963. doi:10.1037/a0031409

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. doi: 10.3758/BRM.41.4.1149

Faust, D., Hart, K. J., & Guilmette, T. J. (1988). Pediatric malingering: The capacity of children to fake believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 56,* 578-582. doi:10.1037/0022-006X.56.4.578

Faust, D., Hart, K. J., Guilmette, T. J., & Arkes, H. R. (1988). Neuropsychologists' capacity to detect adolescent malingerers. *Professional Psychology: Research and Practice, 19,* 508-515. doi:10.1037/0735-7028.19.5.508

Feldman, R. S., Jenkins, L., & Popoola, O. (1979). Detection of deception in adults and children via facial expressions. *Child Development, 50,* 350-355. doi:10.2307/1129409

Fleming, J. G. (1998). *The law of torts* (9th ed.). Sydney, Australia: Thomson Professional.

Frank, M. G., & Ekman, P. (1997). The ability to detect deceit generalizes across different types of high-stake lies. *Journal of Personality and Social Psychology, 72,* 1429-1439. doi:10.1037/0022-3514.72.6.1429

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137,* 201-225. doi:10.1037/0096-3445.137.2.201

Gast, J., & Hart, K. J. (2010). The performance of juvenile offenders on the test of memory malingering. *Journal of Forensic Psychology Practice, 10,* 53-68. doi:10.1080/15228930903173062

Geffen, G. M., Butterworth, P., & Geffen, L. B. (1994). Test-retest reliability of a new form of the Auditory Verbal Learning Test (AVLT). *Archives of Clinical Neuropsychology, 9,* 303-316. doi:10.1016/0887-6177(94)90018-3

Gidley-Larson, J. C., Flaro, L., Peterson, R. L., Connery, A. K., Baker, D. A., & Kirkwood, M. W. (2015). The Medical Symptom Validity Test measures effort not ability in children: A comparison between mild TBI and Fetal Alcohol Spectrum Disorder samples. *Archives of Clinical Neuropsychology, 30,* 192-199. doi:10.1093/arclin/acv012

Golden, C. J. (1978). *Stroop color and word test: A manual for clinical and experimental uses.* Chicago, IL: Stoelting Company.

Gombos, V. A. (2006). The cognition of deception: The role of executive processes in producing lies. *Genetic, Social, and General Psychology Monographs, 132,* 197-214. doi:10.3200/MONO.132.3.197-214

Green, P. (2003). *Medical Symptom Validity Test (MSVT) for Microsoft Windows*: *User's manual and program.* Edmonton, Canada: Author.

Green, P., Flaro, L., Brockhaus, R., & Montijo, J. (2012). Performance on the WMT, MSVT, and NV-MSVT in children with developmental disabilities and in adults with mild traumatic brain injury. In C. R. Reynolds & A. J. Horton (Eds.),

Detection of malingering during head injury litigation (2nd ed., pp. 201-219). New York, NY: Springer Science + Business Media.

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6,* 218–224. doi:10.1037/1040-3590.6.3.218

Gunn, D., Batchelor, J., & Jones, M. (2010). Detection of simulated memory impairment in 6- to 11-year-old children. *Child Neuropsychology, 16,* 105-118. doi:10.1080/09297040903352564

Guttentag, R. E., & Haith, M. M. (1978). Automatic processing as a function of age and reading ability. *Child Development, 49,* 707-716. doi:10.2307/1128239

Harrison, A. G., Flaro, L., & Armstrong, I. (2015). Rates of effort test failure in children with ADHD: An exploratory study. *Applied Neuropsychology: Child*, *4*(3), 197-210. doi:10.1080/21622965.2013.850581

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23,* 1093-1129. doi:10.1080/13854040903155063

Hill, S. K., Ryan, L. M., Kennedy, C. H., & Malamut, B. L. (2003). The relationship between measures of declarative memory and the test of memory malingering in patients with and without temporal lobe dysfunction. *Journal of Forensic Neuropsychology, 3*(3), 1-18. doi:10.1300/J151v03n03_01

Hinz v. Berry, 2 QB 40 (1970). Retrieved from Canlii.

Holmes, O. W., Sr. (1872). The poet at the breakfast table. *The Atlantic Monthly, 29,* p. 231.

Jacobsen, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19. doi:10.1037/0022-006X.59.1.12

Kirk, J. W., Harris, B., Hutaff-Lee, C. F., Koelemay, S. W., Dinkins, J. P., & Kirkwood, M. W. (2011). Performance on the Test of Memory Malingering (TOMM) among a large clinic-referred pediatric sample. *Child Neuropsychology, 17,* 242-254. doi:10.1080/09297049.2010.533166

Kirk, J. W., Hutaff-Lee, C. F., Connery, A. K., Baker, D. A., & Kirkwood, M. W. (2014). The relationship between the self-report BASC-2 validity indicators and performance validity test failure after pediatric mild traumatic brain injury. *Assessment, 21,* 562-569. doi:10.1177/1073191114520626

Kirkwood, M. W. (2012). Overview of tests and techniques to detect negative response bias in children. In E. M. S. Sherman & B. L. Brooks (Eds.), *Pediatric forensic neuropsychology* (pp. 136-161). New York, NY: Oxford University Press.

Kirkwood, M. W. (2015a). A rationale for performance validity testing in child and adolescent assessment. In M. W. Kirkwood (Ed.), *Validity testing in child and adolescent assessment: Evaluating exaggeration, feigning, and noncredible effort* (pp. 3-21). New York, NY: Guilford Press.

Kirkwood, M. W. (2015b). Review of pediatric performance and symptom validity tests. In M. W. Kirkwood (Ed.), *Validity testing in child and adolescent assessment: Evaluating exaggeration, feigning, and noncredible effort* (pp. 79-106). New York, NY: Guilford Press.

Kirkwood, M. W., & Kirk, J. W. (2010). The base rate of suboptimal effort in a pediatric mild TBI sample: Performance on the Medical Symptom Validity Test. *The Clinical Neuropsychologist, 24,* 860-872. doi:10.1080/13854040903527287

Kirkwood, M. W., Yeates, K. O., Randolph, C., & Kirk, J. W. (2012). The implications of symptom validity test failure for ability-based test performance in a pediatric sample. *Psychological Assessment, 24,* 36-45. doi:10.1037/a0024628

Klenberg, L., Korkman, M., & Lahti-Nuuttila, P. (2001). Differential development of attention and executive functions in 3- to 12-year-old Finnish children. *Developmental Neuropsychology, 20,* 407-428. doi:10.1207/S15326942DN2001_6

Koch, W., Douglas, K., Nicholls, T., & O'Neil, M. (2006). *Psychological injuries: Forensic assessment, treatment, and law.* New York, NY: Oxford University Press.

Kochanska, G., Murray, K., Jacques, T. Y., Koenig, A. L., & Vandegeest, K. A. (1996). Inhibitory control in young children and its role in emerging internalization. *Child Development, 67,* 490-507. doi:10.2307/1131828

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: A developmental neuropsychological assessment*. San Antonio, TX: The Psychological Corporation.

Kotai v. Queen of the North (Ship), BCSC 1405 (2009). Retrieved from http://canlii.ca/t/2fksh

Kraut, R.E. (1980). Humans as lie detectors: Some second thoughts. *Journal of Communication, 30,* 209-218. doi:0.1111/j.1460-2466.1980.tb02030.x

Larochette, A.-C., & Harrison, A. G. (2012). Word Memory Test performance in Canadian adolescents with learning disabilities: A preliminary study. *Applied Neuropsychology: Child, 1,* 38-47. doi:10.1080/21622965.2012.665777

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17,* 410-425. doi:10.1076/clin.17.3.410.18089

Larrabee, G. J. (2005). Assessment of malingering. In G. J. Larrabee (Ed.), *Forensic neuropsychology: A scientific approach* (pp. 115-158). New York, NY: Oxford University Press.

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist, 22,* 666-679. doi:10.1080/13854040701494987

Leach, A., Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). 'Intuitive' lie detection of children's deception by law enforcement officials and university students. *Law and Human Behavior, 28,* 661-685. doi:10.1007/s10979-004-0793-0

Lewis, M., Stanger, C., & Sullivan, M. (1989). Deception in 3-year-olds. *Developmental Psychology, 25,* 439-443. doi:10.1037/0012-1649.25.3.439

Libow, J. A. (2000). Child and adolescent illness falsification. *Pediatrics, 105,* 336-342. doi:10.1542/peds.105.2.336

Linden, A. M., & Feldthusen, B. (2015). *Canadian tort law* (10th ed.). Toronto, Canada: LexisNexis.

Lu, P. H., & Boone, K. B. (2002). Suspect cognitive symptoms in a 9-year-old child: Malingering by proxy? *The Clinical Neuropsychologist, 16,* 90-96. doi:10.1076/clin.16.1.90.8328

Lu, Y., Fang J.Q., Tian, L. & Jin, H. (2015). *Advanced Medical Statistics*. Singapore: World Scientific Publishing.

MacAllister, W. S., Nakhutina, L., Bender, H. A., Karantzoulis, S., & Carlson, C. (2009). Assessing effort during neuropsychological evaluation with the TOMM in children and adolescents with epilepsy. *Child Neuropsychology, 15,* 521-531. doi:10.1080/09297040902748226

Macdonald, J. A., Beauchamp, M. H., Crigan, J. A., & Anderson, P. J. (2014). Age-related differences in inhibitory control in the early school years. *Child Neuropsychology, 20,* 509-526. doi:10.1080/09297049.2013.822060

MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109,* 163-203. doi:10.1037/0033-2909.109.2.163

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. *Clinical Neuropsychologist, 29,* 741-776. doi:10.1080/13854046.2015.1087597

Mason v. Westside Cemeteries Ltd., ON SC 8113 (1996). Retrieved from http://canlii.ca/t/1w5ht

McDermott v. Ramadanovic, BC SC 2840 (1988). Retrieved from http://canlii.ca/t/20xhg

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 24,* 1094-1102. doi:10.1076/jcen.24.8.1094.8379

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100. doi:10.1006/cogp.1999.0734

Müller, U., & Kerns, K. (2015). The development of executive function. In L. S. Liben, U. Müller, & R. M. Lerner (Eds.), *Handbook of child psychology and developmental science: Cognitive processes* (7th ed., Vol. 2, pp. 571-623). Hoboken, NJ: John Wiley & Sons.

Mustapha v. Culligan of Canada Ltd., SCC 27 (2008). Retrieved from http://canlii.ca/t/1wz6f

Nagle, A. M., Everhart, D. E., Durham, T. W., McCammon, S. L., & Walker, M. (2006). Deception strategies in children: Examination of forced choice recognition and verbal learning and memory techniques. *Archives of Clinical Neuropsychology, 21,* 777-785. doi:10.1016/j.acn.2006.06.011

Newton, P., Reddy, V., & Bull, R. (2000). Children's everyday deception and performance on false-belief tasks. *British Journal of Developmental Psychology, 18,* 297-317. doi:10.1348/026151000165706

Odhavji Estate v. Woodhouse, SCC 69 (2003). Retrieved from http://canlii.ca/t/1g18n

Oldershaw, L., & Bagby, R. (1997). Children and deception. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 153-166). New York, NY: Guilford Press.

Osborne, P. H. (2015). *The law of torts* (5th ed.). Toronto, Canada: Irwin Law.

O'Sullivan, M. (2007). Unicorns or Tiger Woods: Are lie detection experts myths or rarities? A response to On lie detection 'Wizards' by Bond and Uysal. *Law and Human Behavior, 31,* 117-123. doi:10.1007/s10979-006-9058-4

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior, 33,* 530-538. doi:10.1007/s10979-008-9166-4

Perna, R., & Loughan, A. R. (2014). The influence of effort on neuropsychological performance in children: Is performance on the TOMM indicative of neuropsychological ability? *Applied Neuropsychology: Child, 3,* 31-37. doi:10.1080/21622965.2012.686339

Petersen, I. T., Hoyniak, C. P., McQuillan, M. E., Bates, J. E., & Staples, A. D. (2016). Measuring the development of inhibitory control: The challenge of heterotypic continuity. *Developmental Review, 40, 25*-71. doi:10.1016/j.dr.2016.02.001

Podell, K., DeFina, P.A., Barrett, P., McCullen, A.M., & Goldberg, E, 2003). Assessment of Neuropsychological Functioning. In I. Weiner (Ed.), *Handbook of Psychology: Assessment Psychology.* New Jersey: John Wiley & Sons.

Polak, A., & Harris, P. (1999). Deception by young children following noncompliance. *Developmental Psychology, 35,* 561-568. doi:10.1037/0012-1649.35.2.561

Psychological Corporation. (2001). *Wechsler Individual Achievement Test* (Abbreviated, 2nd ed.). San Antonio, TX: Author.

R. v. Béland, 2 SCR 398 SCC (1987). Retrieved from http://canlii.ca/t/1ftm1

Rabbitt, P. (Ed.). (1998). *Methodology of frontal and executive function*. Sussex, UK: Psychology Press.

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment, 10,* 10-20. doi:10.1037/1040-3590.10.1.10

Rey, A. (1958). *L'examen clinique en psychologie* [Clinical assessment in psychology]. Paris, France: Presses Universitaires.

Rienstra, A., Spaan, P. J., & Schmand, B. (2010). Validation of symptom validity tests using a 'child-model' of adult cognitive impairments. *Archives of Clinical Neuropsychology, 25,* 371-382. doi:10.1093/arclin/acq035

Rogers, R. (Ed.) (1997). *Clinical assessment of malingering and deception* (2nd ed.). New York, NY: Guilford Press.

Saadati v. Moorhead, BCCA 393 (2015). Retrieved from http://canlii.ca/t/gl8hk

Salekin, R., Kubak, F., & Lee, Z. (2008). Deception in children and adolescents. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (3rd ed., pp. 343-364). New York, NY: Guilford Press.

Schneider, H. E., Kirk, J. W., & Mahone, E. M. (2014). Utility of the Test of Memory Malingering (TOMM) in children ages 4-7 years with and without ADHD. *The Clinical Neuropsychologist, 28,* 1133-1145. doi:10.1080/13854046.2014.960004

Sharland, M. J., & Gfeller, J. D. (2007). A survey of neuropsychologists' beliefs and practices with respect to the assessment of effort. *Archives of Clinical Neuropsychology, 22,* 213-223. doi:10.1016/j.acn.2006.12.004

Sherman, E. M. S., & Brooks, B. L. (Eds.). (2012). *Pediatric forensic neuropsychology*. New York, NY: Oxford University Press.

Slick, D. J., & Sherman, E. M. S. (2012). Differential diagnosis of malingering and related clinical presentations. In E. M. S. Sherman & B. L. Brooks (Eds.), *Pediatric forensic neuropsychology* (pp. 113-135). New York, NY: Oxford University Press.

Slick, D. J., & Sherman, E. M. S. (2013). Differential diagnosis of malingering. In D. A. Carone & S. S. Bush (Eds.), *Mild traumatic brain injury: Symptom validity assessment and malingering* (pp. 57-72). New York, NY: Springer.

Slick, D. J., Sherman, E., & Iverson, G. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *Clinical Neuropsychologist, 13*, 545-561. doi:10.1076/1385-4046(199911)13:04;1-Y;FT545

Slick, D. J., Tan, J. E., Sherman, E., & Strauss, E. H. (2010). Malingering and related conditions in pediatric populations. In A.S. Davis (Ed.), *Handbook of pediatric neuropsychology* (pp. 457-470). New York, NY: Springer.

Slick, D. J., Tan, J. E., Strauss, E. H., & Hultsch, D. F. (2004). Detecting malingering: A survey of experts' practices. *Archives of Clinical Neuropsychology, 19,* 465-473. doi:10.1016/j.acn.2003.04.001

Strauss, E. H., Sherman, E., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York, NY: Oxford University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643-662. doi:10.1037/h0054651

Talwar, V., & Lee, K. (2002a). Development of lying to conceal a transgression: Children's control of expressive behaviour during verbal deception. *International Journal of Behavioral Development, 26,* 436-444. doi:10.1080/01650250143000373

Talwar, V., & Lee, K. (2002b). Emergence of white-lie telling in children between 3 and 7 years of age. *Journal of Developmental Psychology, 48,* 160-181. doi:10.1353/mpq.2002.0009

Talwar, V., & Lee, K. (2008). Social and cognitive correlates of children's lying behavior. *Child Development, 79,* 866-881. doi:10.1111/j.1467-8624.2008.01164.x

Talwar, V., Gordon, H., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology, 43,* 804-810. doi:10.1037/0012-1649.43.3.804

Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). Children's lie-telling to conceal a parent's transgression: Legal implications. *Law & Human Behavior, 28,* 411-435. doi:10.1023/B:LAHU.0000039333.51399.f6

Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2006). Adults' judgments of children's coached reports. *Law and Human Behavior, 30,* 561-570. doi:10.1007/s10979-006-9038-8

Talwar, V., Murphy, S., & Lee, K. (2007). White lie-telling in children for politeness purposes. *International Journal of Behavioral Development, 31,* 1-11. doi:10.1177/0165025406073530

Talwar, V., Renaud, S., & Conway, L. (2015). Detecting children's lies: Are parents accurate judges of their own children's lies? *Journal of Moral Education, 44,* 81-96. doi:10.1080/03057240.2014.1002459

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment, 9,* 260-268. doi:10.1037/1040-3590.9.3.260

Tye, M. C., Amato, S. L., Honts, C. R., Devitt, M. K., & Peters, D. (1999). The willingness of children to lie and the assessment of credibility in an ecologically relevant laboratory setting. *Applied Developmental Science, 3,* 92-109. doi:10.1207/s1532480xads0302_4

Undeutsch, U. (1989). The development of statement reality analysis. In J. C. Yuille (Ed.), *Credibility assessment* (pp. 101-119). New York, NY: Kluwer Academic/Plenum.

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence, 35,* 427-449. doi:10.1016/j.intell.2006.09.001

Vakil, E., Blachstein, H., & Sheinman, M. (1998). Rey AVLT: Developmental norms for children and the sensitivity of different memory measures to age. *Child Neuropsychology, 4,* 161-177. doi:10.1076/chin.4.3.161.3173

Vanek v. Great Atlantic & Pacific Company of Canada Limited, ON CA 2863 (1999). Retrieved from http://canlii.ca/t/1f9ws

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and the implications for professional practice.* Chichester, UK: Wiley.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. *Psychology, Public Policy, and Law, 11,* 3-41. doi:10.1037/1076-8971.11.1.3

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2002). Will the truth come out? The effect of deception, age, status, coaching, and social skills on CBCA scores. *Law and Human Behavior, 26,* 261-283. doi:10.1023/A:1015313120905

Vrij, A. & Fisher, R. (2016). Which lie detection tools are ready for use in the criminal justice system? *Journal of Applied Research in Memory and Cognition, 5*, 302-307. doi.org/10.1016/j.jarmac.2016.06.014

Walker, J. S. (2011). Malingering in children: Fibs and faking. *Child and Adolescent Psychiatric Clinics of North America, 20,* 547-556. doi:10.1016/j.chc.2011.03.013

Wechsler, D., Kaplan, E., Fein, D., Kramer, J., Morris, R., Delis, D., & Maelender, A. (2003). *Wechsler Intelligence Scale for Children: Fourth edition (WISC-IV)*. San Antonio, TX: Pearson.

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. New York, NY: Oxford University Press.

Welsh, M., & Pennington, B. (1988). Assessing frontal lobe functioning in children: Views from developmental psychology. *Developmental Neuropsychology, 4,* 199-230. doi:10.1080/87565648809540405

Welsh, M., Pennington, B., & Groisser, D. (1991). A normative-developmental study of executive function: A window on prefrontal function in children. *Developmental Neuropsychology, 7,* 131-149. doi:10.1080/87565649109540483

Westcott, H., Davies, G., & Clifford, B. (1991). Adults' perceptions of children's videotaped truthful and deceptive statements. *Children & Society, 5,* 123-135. doi:10.1111/j.1099-0860.1991.tb00378.x

Williams, S., Moore, K., Crossman, A. M., & Talwar, V. (2016). The role of executive functions and theory of mind in children's prosocial lie-telling. *Journal of Experimental Child Psychology*, *141,* 256-266. doi:10.1016/j.jecp.2015.08.001

Wilson, A., Smith, M., & Ross, H. (2003). The nature and effects of young children's lies. *Social Development, 12,* 21-45. doi:10.1111/1467-9507.00220

Woodward, M. (2005). *Epidemiology: Study design and data analysis.* London, UK: Chapman and Hall/CRC.

Wright, I., Waterman, M., Prescott, H., & Murdoch-Eaton, D. (2003). A new Stroop-like measure of inhibitory function development: Typical developmental trends.

Journal of Child Psychology and Psychiatry, 44, 561-575. doi:10.1111/1469-7610.00145

Yoshikawa v. Yu. 21 BCAC (1996). Retrieved from QuickLaw.

Zelazo, P. D., & Müller, U. (2002). Executive function in typical and atypical development. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 445-469). Malden, MA: Blackwell.

Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of General Psychology, 1,* 198-226. doi:10.1037/1089-2680.1.2.198

Appendix A.

Medical History Questionnaire

SIMON FRASER UNIVERSITY

Medical History Questionnaire

Child's Name: _____ Gender: _____ Age: _____ Grade: _____

(please circle)

1) Does your child require corrective lenses or glasses? YES NO
2) Does your child have difficulty with hearing? YES NO
3) Is your child colour-blind? YES NO
4) Has your child ever seen a health/mental health profession for any of the following problems:
   a. Head injury / Concussion YES NO
   b. Learning disability, such as dyslexia YES NO
   c. Neurological problems, such as epilepsy YES NO

   Would any of these conditions affect their performance during the testing session in this study? YES NO
5) Has your child ever been diagnosed with Attention Deficit Disorder (ADD) or Attention Deficit Hyperactivity Disorder (ADHD)? YES NO
   Is it currently being managed? YES NO
6) Does your child have any other chronic or acute conditions that may interfere with their ability to participate in this research? YES NO

a. If yes, please specify:

_____

Appendix B.

Script for Obtaining Assent and Debriefing Participants

Script for Obtaining Assent

"Today, some children in your class are going to be playing some games where they have to remember words, solve puzzles, and answer questions. You don't have to do it and we can stop or take breaks at any time. If you don't want to play the games any more, I can take you back to your classroom. Do you have any questions? Would you like to participate?"

Script for Debriefing

Best Effort Condition: "Today we were playing games where you had to try your best. You did a great job! Because you did so well, I have a special prize for you."

Withholding Condition: "Today we were playing games where you had to pretend that you were like the child in the storybook – it was like other pretend games where you act like somebody else. It's important to remember that this was just a game and only for today. When you go back to class, the game is over and if your teacher or your parents ask you to try your best at something, it's important that you listen to them and try your best. Do you have any questions about this? Okay, you did an excellent job today and earned a special prize!"

Appendix C.

Counterbalanced Test Order

P#	Instr	Study 1: Session 1 Test Order						
1	C	RAVLT(A)	Nepsy	Coding	SS	Info	Stroop	WIAT
2	W	Nepsy	RAVLT(A)	SS	Stroop	Coding	Info	WIAT
3	C	Nepsy	WIAT	RAVLT(A)	Info	Coding	SS	Stroop
4	W	RAVLT(A)	SS	Coding	Nepsy	Info	Stroop	WIAT
5	C	Info	RAVLT(A)	Coding	SS	Stroop	WIAT	Nepsy
6	W	SS	Info	RAVLT(A)	WIAT	Coding	Nepsy	Stroop
7	C	RAVLT(A)	Nepsy	SS	Info	Coding	Stroop	WIAT
8	W	SS	RAVLT(A)	Info	WIAT	Nepsy	Stroop	Coding
9	C	Coding	Nepsy	RAVLT(A)	WIAT	SS	Stroop	Info
10	W	RAVLT(C)	Stroop	Coding	SS	WIAT	Info	Nepsy
11	C	WIAT	RAVLT(A)	SS	NEPSY	Info	Stroop	Coding
12	W	Stroop	Info	SS	RAVLT(C)	WIAT	NEPSY	Coding
13	C	NEPSY	Stroop	Coding	RAVLT(A)	SS	Info	WIAT
14	W	Info	WIAT	Coding	NEPSY	Stroop	RAVLT(C)	SS

15	C	SS	Coding	RAVLT(A)	Stroop	NEPSY	WIAT	Info
16	W	SS	Info	Stroop	NEPSY	C	RAVLT(C)	WIAT
17	C	Stroop	WIAT	Coding	RAVLT(A)	Info	NEPSY	SS
18	W	RAVLT(C)	WIAT	Stroop	SS	Coding	Info	NEPSY
19	C	WIAT	Info	SS	NEPSY	RAVLT(A)	Coding	Stroop
20	W	NEPSY	RAVLT(C)`	Info	Coding	WIAT	Stroop	SS
21	C	WIAT	SS	Coding	Stroop	Info	NEPSY	RAVLT(A)
22	W	SS	WIAT	Info	Stroop	Coding	NEPSY	RAVLT(C)
23	C	Coding	Info	NEPSY	Stroop	WIAT	SS	RAVLT(A)
24	W	Stroop	NEPSY	RAVLT(C)	WIAT	Coding	Info	SS
25	C	Info	SS	NEPSY	Stroop	RAVLT(A)	Coding	WIAT
26	W	RAVLT(C)	SS	Coding	NEPSY	Info	WIAT	Stroop
27	C	Stroop	SS	WIAT	RAVLT(A)	Coding	NEPSY	Info
28	W	SS	WIAT	NEPSY	Info	Coding	Stroop	RAVLT(C)
29	C	Coding	NEPSY	RAVLT(A)	Info	Stroop	SS	WIAT
30	W	RAVLT(C)	SS	Stroop	WIAT	NEPSY	Info	Coding
31	C	RAVLT(A)	Info	NEPSY	SS	WIAT	Stroop	Coding

32	W	WIAT	RAVLT(C)	Info	Stroop	Coding	NEPSY	SS
33	C	RAVLT(A)	Info	SS	Coding	WIAT	Stroop	NEPSY
34	W	Stroop	Coding	Info	SS	WIAT	RAVLT(A)	NEPSY
35	C	WIAT	RAVLT(C)	NEPSY	Info	SS	Stroop	CODING
36	W	Info	Stroop	SS	Coding	RAVLT(A)	WIAT	NEPSY
37	C	Info	Coding	RAVLT(C)	WIAT	Stroop	NEPSY	SS
38	W	Coding	NEPSY	Info	Stroop	WIAT	SS	RAVLT(A)
39	C	Coding	SS	Stroop	WIAT	NEPSY	Info	RAVLT(C)
40	W	NEPSY	WIAT	Info	Coding	RAVLT(A)	Stroop	SS
41	C	Info	RAVLT(C)	WIAT	Stroop	NEPSY	SS	Coding
42	W	WIAT	RAVLT(A)	SS	Stroop	Coding	Info	NEPSY
43	C	Stroop	Coding	SS	NEPSY	Info	WIAT	RAVLT(C)
44	W	Info	RAVLT(A)	SS	NEPSY	Coding	Stroop	WIAT
45	C	SS	Stroop	WIAT	NEPSY	Info	Coding	RAVLT(C)
46	W	SS	WIAT	Coding	RAVLT(A)	NEPSY	Stroop	Info
47	C	NEPSY	RAVLT(C)	Coding	Stroop	WIAT	Info	SS
48	W	RAVLT(A)	SS	Info	Stroop	Coding	WIAT	NEPSY

49	C	NEPSY	Stroop	Coding	RAVLT(C)	Info	SS	WIAT
50	W	Info	WIAT	NEPSY	Coding	Stroop	SS	RAVLT(A)
51	C	Info	NEPSY	Stroop	RAVLT(C)	WIAT	SS	Coding
52	W	SS	WIAT	Stroop	RAVLT(A)	Coding	NEPSY	Info
53	C	Info	Stroop	RAVLT(C)	Coding	NEPSY	SS	WIAT
54	W	SS	Stroop	Coding	Info	NEPSY	RAVLT(A)	WIAT
55	C	Info	WIAT	RAVLT(C)	NEPSY	Coding	Stroop	SS
56	W	WIAT	Stroop	RAVLT(A)	Info	NEPSY	SS	CODING
57	C	NEPSY	Stroop	WIAT	RAVLT(C)	SS	Coding	Info
58	W	Coding	SS	Info	Stroop	RAVLT(A)	NEPSY	WIAT
59	C	Coding	RAVLT(C)	SS	NEPSY	Info	WIAT	Stroop
60	W	Coding	Info	NEPSY	SS	Stroop	WIAT	RAVLT(A)

P#	Instruction	Study 1: Session 2 Test Order				
1	C	MSVT	SS	Coding	MSVT-Delay	RAVLT*
2	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
3	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
4	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
5	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
6	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
7	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
8	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
9	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
10	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
11	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
12	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
13	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
14	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
15	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
16	W	RAVLT	MSVT	SS	Coding	MSVT-Delay

17	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
18	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
19	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
20	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
21	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
22	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
23	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
24	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
25	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
26	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
27	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
28	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
29	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
30	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
31	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
32	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
33	C	RAVLT	MSVT	SS	Coding	MSVT-Delay

34	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
35	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
36	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
37	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
38	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
39	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
40	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
41	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
42	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
43	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
44	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
45	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
46	W	MSVT	SS	Coding	MSVT-Delay	RAVLT
47	C	MSVT	Coding	SS	MSVT-Delay	RAVLT
48	W	MSVT	Coding	SS	MSVT-Delay	RAVLT
49	C	MSVT	SS	Coding	MSVT-Delay	RAVLT
50	W	MSVT	SS	Coding	MSVT-Delay	RAVLT

51	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
52	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
53	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
54	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
55	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
56	W	RAVLT	MSVT	SS	Coding	MSVT-Delay
57	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
58	W	RAVLT	MSVT	SS	Info	MSVT-Delay
59	C	RAVLT	MSVT	SS	Coding	MSVT-Delay
60	W	RAVLT	MSVT	SS	Coding	MSVT-Delay

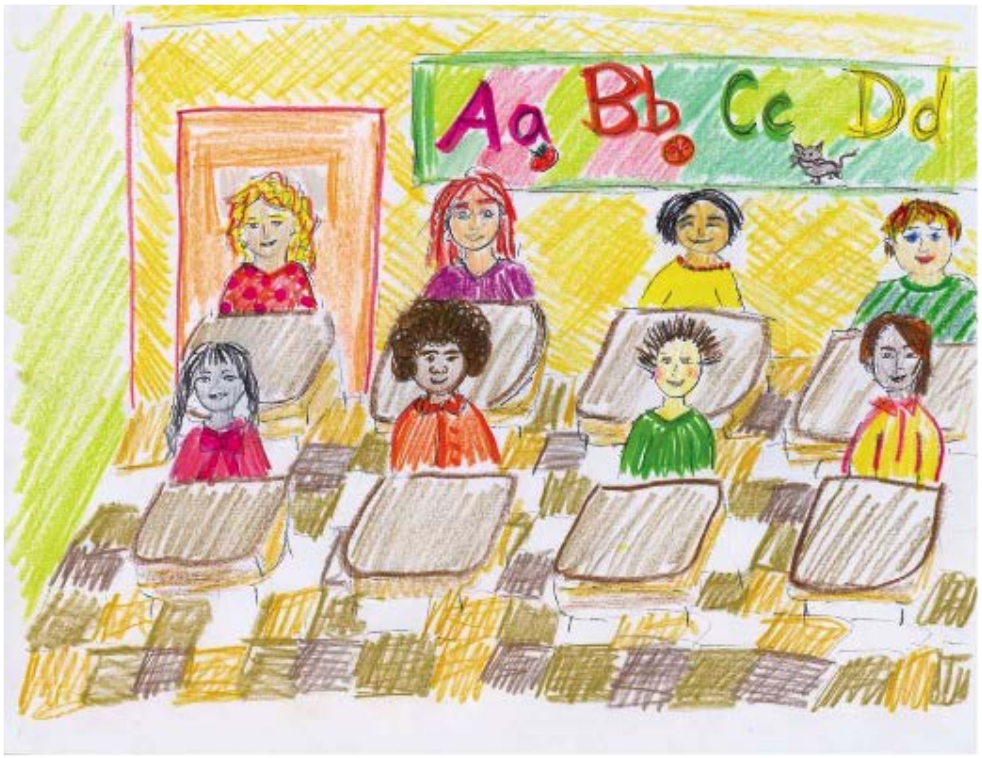
* Version of RAVLT administered was opposite to that given to the participant in testing session #1

Appendix D.

Storybooks

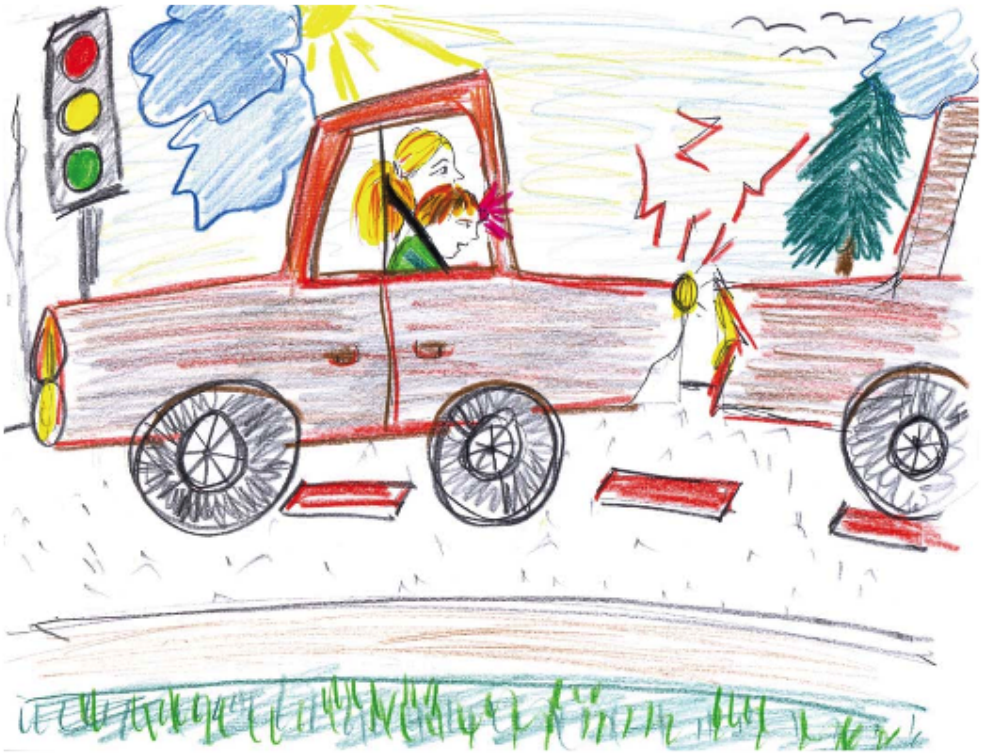
Storybook for Participants in Withholding Condition

Jack and his Bump on the Head



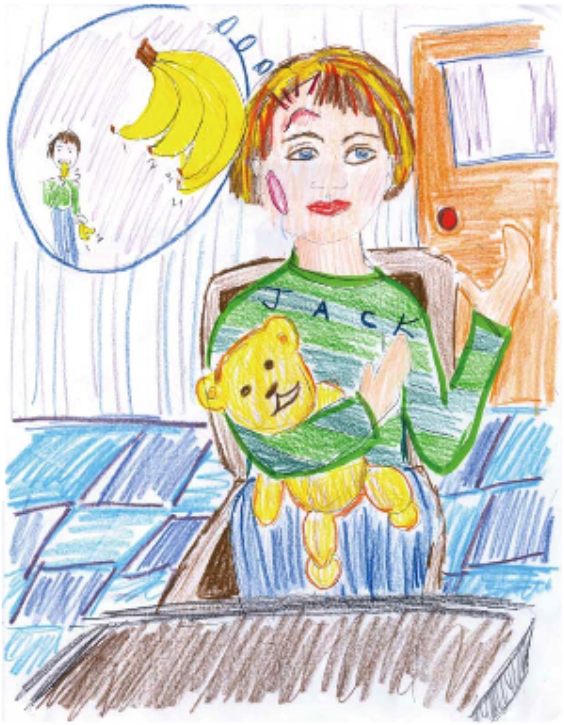
This story is about a smart kid named Jack who goes to school, works hard, likes playing games at recess, and has lots of friends.

Page 1



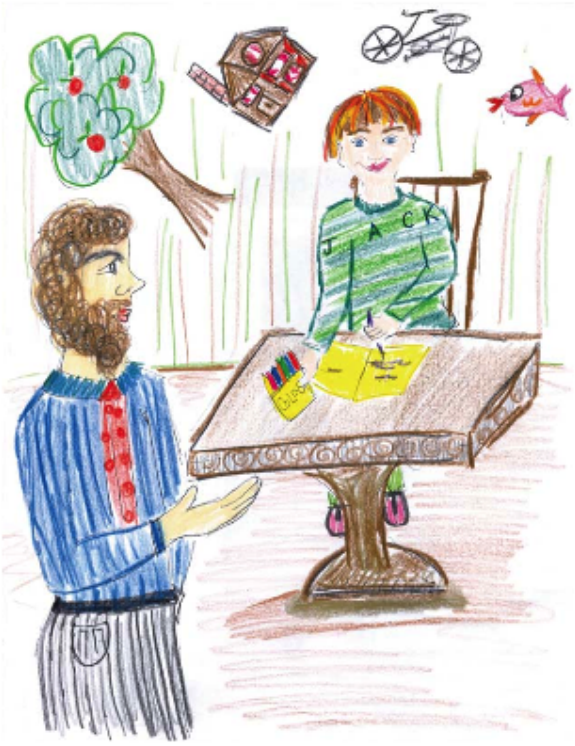
One day, he was driving home with his mom and they got into a car accident. In the accident, Jack bumped his head and started acting a little confused. The ambulance came to the accident and said that he's probably okay, but he should have a quick check at the hospital just to make sure. At the hospital, the doctors asked him a few questions and shined a bright flashlight into his eyes to see if he was okay. The doctors said that he had a mild concussion and should see a psychologist next week to see if there's any damage to his head.

Page 2



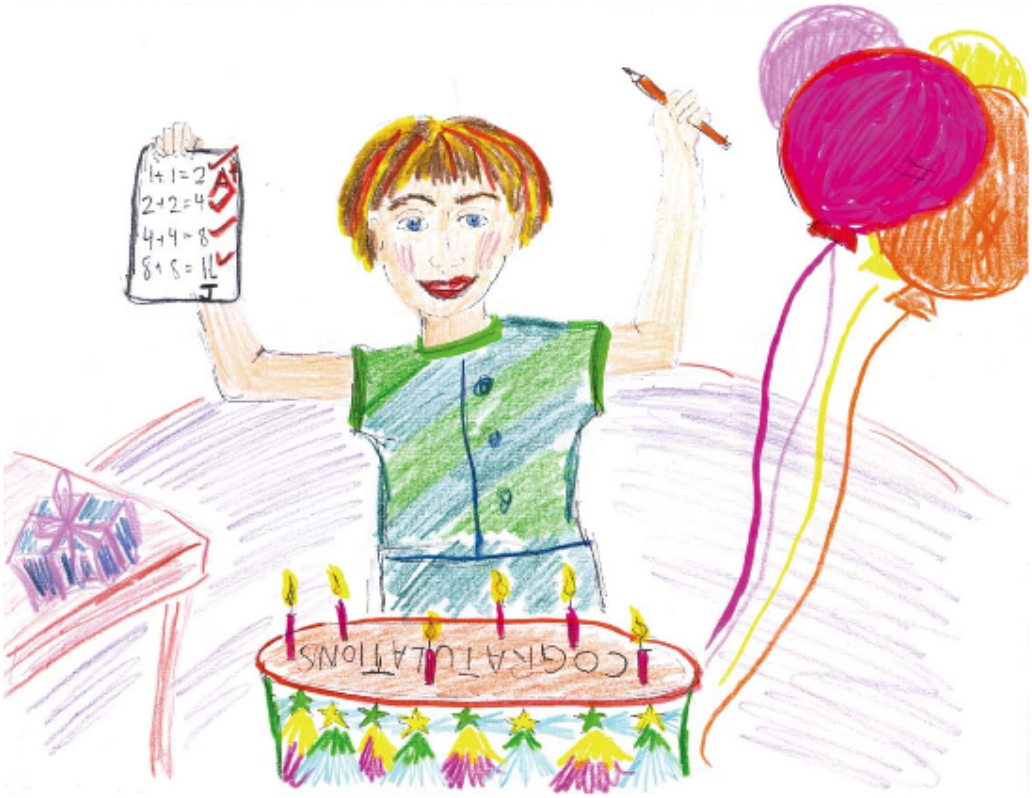
A week later, Jack's mom takes him to see a psychologist. The psychologist explains that most children who bump their head are okay after a few days or a few weeks, but sometimes the injury is more severe and the problems don't go away. The psychologist says that it's not very common, but some children forget things and do things slower than they could have before the accident.

The psychologist then starts asking Jack some questions to test his memory and thinking. He asks, "If I have 4 bananas and eat 3 of them, how many do I have left?" Jack knows the answer to this one and says, "you have 1 left". The psychologist then asks a more difficult question, "If packs of gum cost $1.50 and I buy 3 packs, how much will the gum cost?" Jack thinks for a minute. He knows that he should know the answer because he could answer these types of questions before the accident, but he can't think of the right answer. Jack replies, "maybe $5." The psychologist then says, "I would like you to remember this list of words, House, Tree, Bicycle, Fish". Jack tries hard to remember and can only remember three of the words, "umm, House, Bicycle, and Fish".



The psychologist then asks Jack to copy some pictures as fast as he can. Jack's excited for this test because he used to be pretty fast at drawing, but when he starts the test, he realizes that he can't draw as fast as he could before. The psychologist then asks one final question and says, "remember that list of words I asked you to remember, can you tell me the words again?" Jack responds with the same words he remembered last time, "House, Bicycle, and Fish". Just like last time, he couldn't remember the second word, "Tree".

As the psychologist expected, Jack had a hard time answering difficult questions, he was slower than he was before the accident, and he keeps forgetting the same word. The psychologist tells Jack and his mom that it's okay he didn't do so well because he's still getting better and will be back to normal soon. The psychologist also says that it's good that he wore a seatbelt because it protected him from bumping his head even worse.



After a few weeks, Jack was feeling much better and could do all of the things he could do before the accident, he could answer difficult questions, draw pictures quickly, and remember lots of words. Luckily, the bump wasn't so bad after all.

Parents/Guardians: Please read these instructions to your child at the end of the story. Thank you!

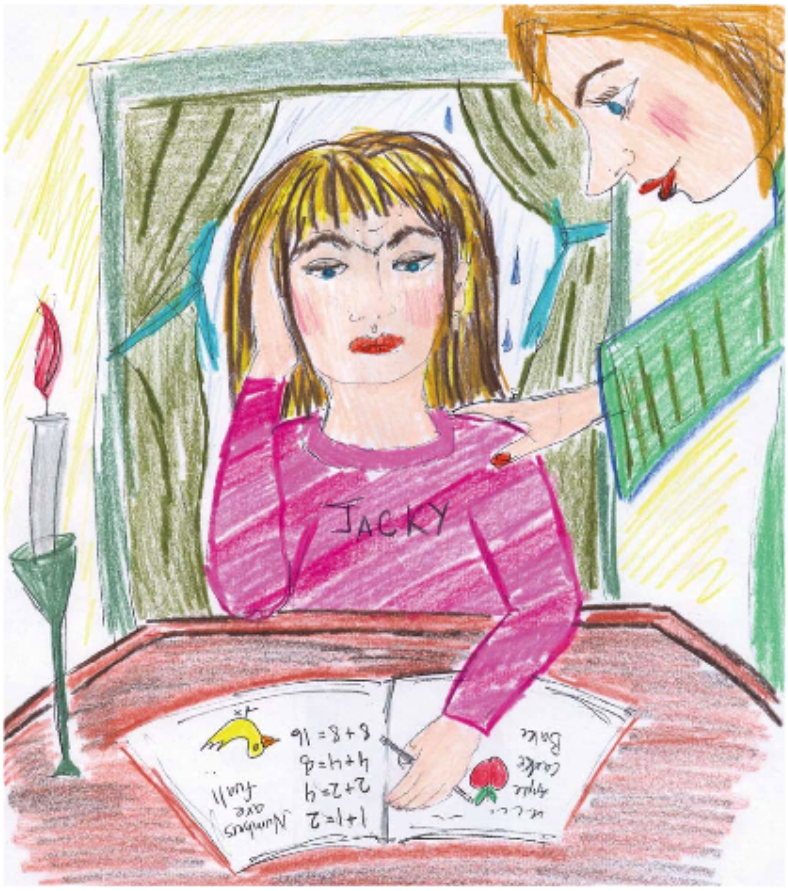
"When you are at school tomorrow, a person from the university will be coming to your class and will ask you to do some puzzles and answer some questions. When you answer the questions, I want you to answer the questions just like Jack. This means you should: (1) Do things slower than you could normally do, (2) pretend that you can't remember as much as you normally could, and (3) if the question seems easy, get it right, but if the question seems difficult, get it wrong."

Storybook for Participants in Best Effort Condition

JACKY AND HER TROUBLE WITH MATH

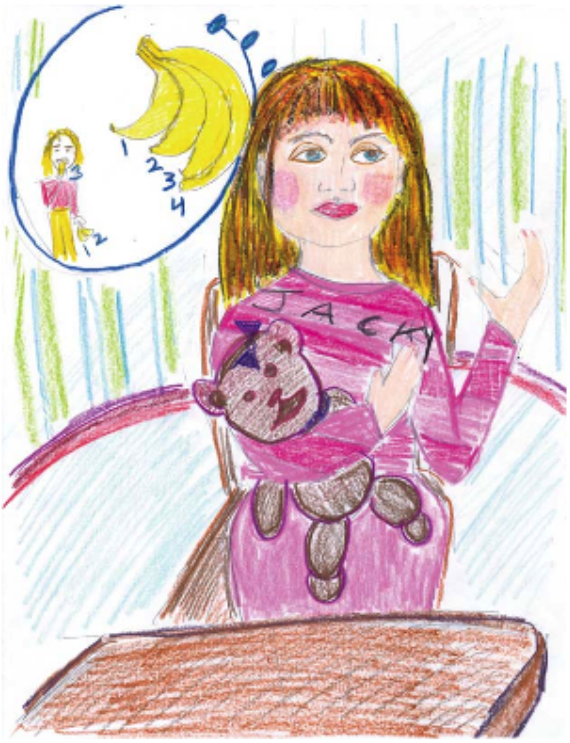
This story is about a smart kid named Jacky who goes to school, works hard, likes playing games at recess, and has lots of friends.

Page 1

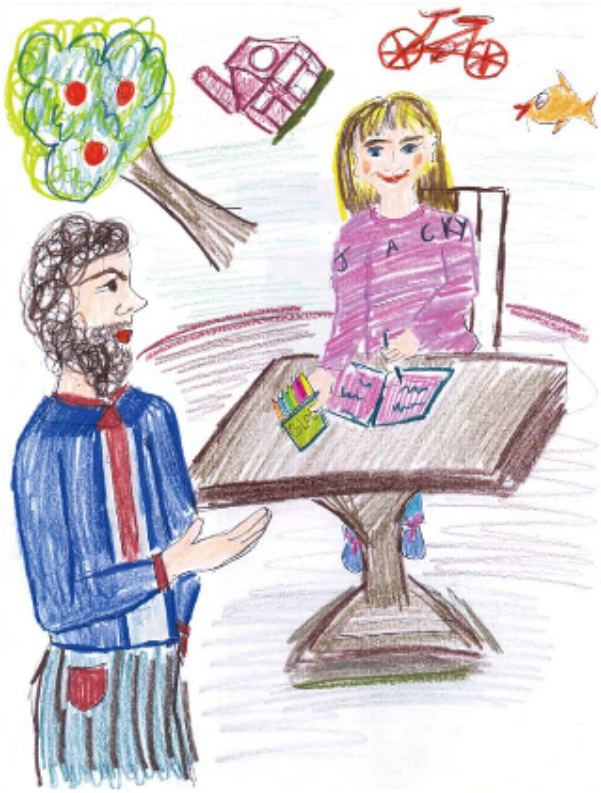


Lately, she's been having a hard time with her math homework. She tries hard to keep up, but she's still confused. She asks her teacher for help and she explains some of the problems so that they become easier, but it seems to take her much longer to learn new things. Her parents also give her some extra help at home, but she still takes longer than all of his classmates to understand new types of math.

Page 2

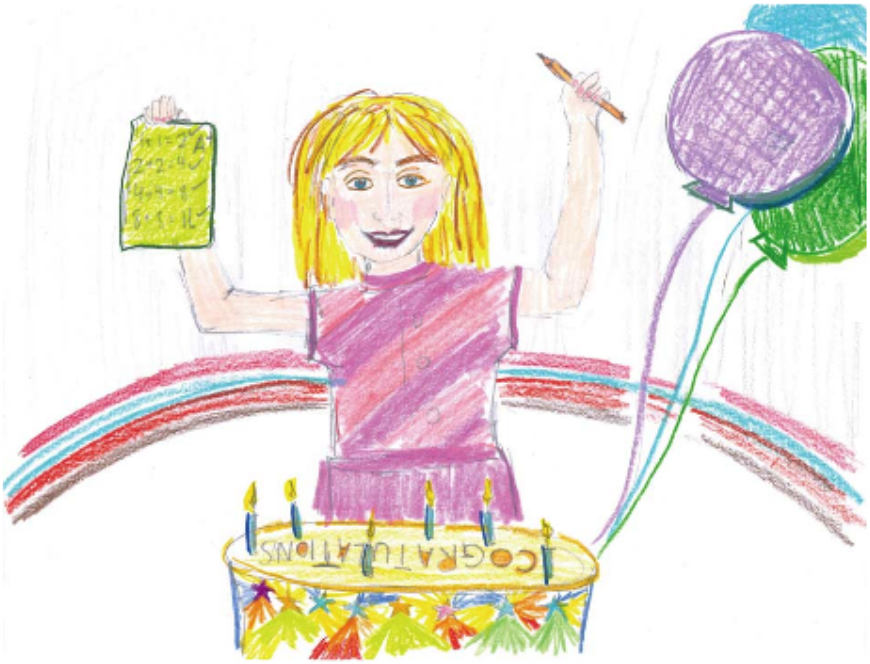


One day, her mom takes her to see a psychologist to find out if there's anything that can help. The psychologist gives Jacky some tests and she tries her best. The psychologist asks, "If I have 4 bananas and I eat 3 of them, how many do I have left?" Jacky replies, "You have 1". "The psychologist then asks a much tougher question, "If packs of gum cost $1.50 and I buy 3 packs, how much will the gum cost?". Jacky doesn't know the answer to this one and replies, "umm, about $5".



The psychologist then says, "Jacky, I would like you to remember this list of words, House, Tree, Bicycle, Fish". Jacky tries hard to remember and replies, "House, Tree, Bicycle, and Fish". The psychologist then asks Jacky to copy some pictures as fast as she can. Jacky's excited for this test because she's pretty fast at drawing. The psychologist then asks one final question and says, "remember that list of words I asked you to remember, can you tell me the words again?" Jacky responds with the same words she remembered last time, "House, Tree, Bicycle, and Fish". Just like last time, she could remember all of the words.

After a while, the psychologist tells Jacky that she's very good at lots of subjects and needs to keep working hard at math. She recommends that Jacky get some extra help after school and to keep trying her best, because the more she practices, the easier it will be. Jacky and her mom thank the psychologist and go home.



After a while of getting extra help, Jacky starts to notice that she's starting to understand more and more of her homework. To her surprise, she even starts getting good marks on her tests and her parents and teachers are proud of her improvement. She learned that even though some subjects are more difficult for her than other ones, she shouldn't be discouraged because sometimes, she just needs a little help and that's okay.

Parents/Guardians: Please read these instructions to your child at the end of the story. Thank you!

"When you are at school tomorrow, a person from the university will be coming to your class and will ask you to do some puzzles and answer some questions, just like you did last week. When you answer the questions, I want you to try your best on all of the questions, even if they seem difficult."