A DISEASE-SPECIFIC HEALTH STATUS MEASUREMENT FOR CHILDREN WITH HYDROCEPHALUS

By

ABHAYA V. KULKARNI, M.D., M.Sc.

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

MEASUREMENT OF HEALTH STATUS IN CHILDREN WITH HYDROCEPHALUS

DOCTOR OF PHILOSOPHY (2003) McMaster University,
(Health Research Methodology) Hamilton, Ontario

TITLE: A Disease-Specific Health Status Measurement for Children with Hydrocephalus

AUTHOR: Abhaya Vivek Kulkarni, M.D. (University of Toronto), M.Sc. (McMaster University)

SUPERVISOR: Professor Michael H. Boyle

NUMBER OF PAGES: viii, 172

Abstract

Hydrocephalus is a common condition of childhood. Attempts to measure the health status of children with hydrocephalus have traditionally relied on surgical outcomes, non-specific generic health outcomes, or very specific neuropsychological measures. This work describes the development of a new disease-specific health status outcome measure for children with hydrocephalus – called the Hydrocephalus Outcome Questionnaire (HOQ).

This work begins with a discussion of several methodological issues relevant to health status measurement, highlighting certain points of controversy. This is followed by a review of the methodology and the results of various stages of development of this new health status measure. This includes the stages of concept development, item generation, item reduction, reliability testing, and validity testing. The final section describes the use of some different approaches to providing interpretability to the new outcome measure. This work was approved by the Research Ethics Board at the Hospital for Sick Children, Toronto.

The result of this work was the 60-item Hydrocephalus Outcome Questionnaire. It demonstrated very good psychometric properties and was well received by the parents of children with hydrocephalus, who are the primary respondents. It is hoped that this will serve a useful role as a much-needed outcome measure for pediatric hydrocephalus.

Acknowledgements

This project has been the culmination of a great deal of work involving many dedicated people. I would like to thank the following individuals for helping out at various stages of this project to make it all possible: James Drake, John Kestle, Maria Lamberti-Pasculli, Doron Rabin, Patti Rowe, and Diane Vitaliano. My thanks and appreciation also to my thesis committee: Michael Boyle (supervisor), Barbara Schmidt, and Paul Stratford.

The following organizations contributed financially towards the successful completion of this project, and, for that, I am very grateful to each of them:

- *Hospital for Sick Children Foundation*, for providing project funding as well as awarding me the Duncan L. Gordon Fellowship for two years.
- *Congress of Neurological Surgeons*, for awarding me the Wilder Penfield Clinical Investigation Fellowship.
- *Physicians' Services Incorporated*, for providing project funding.
- *Bloorview Childrens Hospital Foundation,* for awarding me a graduate student scholarship.

I would also like to thank the nearly 200 children and their parents who took the time to participate in all stages of this project. Their time, their input, and their unwavering enthusiasm was greatly appreciated and made this research so much easier.

A final thanks to my beautiful wife and my parents for all their support and understanding throughout this rather lengthy endeavour.

- AVK, May 2003

Table of Contents

1. Hydrocephalus and the Need for a New Outcome Measure[§]

1.1 Clinical Review of Hydrocephalus

Hydrocephalus is a disorder in which an excess amount of cerebrospinal fluid (CSF) accumulates in the ventricular system of the brain. The primary etiologies vary and children of all ages can be affected.

1.1.1 Epidemiology

The reported incidence of infantile hydrocephalus is approximately 2 to 4 per 1000 live births.[2, 3] It ranks as the second most common congenital neurological malformation in North America, after spina bifida.[4] However, this is probably a gross underestimate of the overall societal disease burden, since many cases of hydrocephalus are associated with other conditions and are not diagnosed until later in life. In children born with spina bifida, for example, nearly 80% develop hydrocephalus secondarily.[5] The annual incidence of newly diagnosed hydrocephalus requiring shunt surgery was estimated at 1 per 12,615 population in Canada.[6] Based on the 1988 United States National Health Interview Survey, it has been estimated that there are 125,000 people with shunts in the United States, with new shunts being inserted at a rate of over 18,000 per year.[7] Based on this same data, it was calculated that shunt-related procedures cost the American health care system US$ 94 million annually, of which over half was devoted to shunt revisions.

1.1.2 Etiology

While there are many etiologies for hydrocephalus, the basic problem relates to an imbalance between the *production* of the CSF by the choroid plexus and the *absorption* of CSF, mostly by the arachnoid villi. This imbalance leads to an accumulation of CSF in the brain, and it is this expansion and increase in intracranial pressure that detrimentally effects brain function and development.

Hydrocephalus resulting from *overproduction* of CSF is exceedingly rare, occurring only in association with choroid plexus papillomas. These are rare tumours and

---

[§] This review chapter is taken largely from the candidate's Masters thesis[1] and is not being presented here as original new work. It is provided simply as background reading for the remainder of this Doctoral thesis.

they account for less than 1% of all cases of hydrocephalus.[8]

Obstruction of CSF flow, thereby preventing its absorption, is, by far, the most common cause of hydrocephalus.[9] This obstruction can occur anywhere from the point of CSF production by the choroid plexus to the point of absorption by the arachnoid villi. The types of obstructions have been traditionally divided into those that result in obstruction within the ventricular system (*non-communicating hydrocephalus*) and those that cause CSF obstruction beyond the ventricular system (*communicating hydrocephalus*). Table 1.1 lists the various causes of obstructive hydrocephalus.

TABLE 1.1 Classification of Hydrocephalus

(from Milhorat, 1996)[9]

NON-COMMUNICATING	COMMUNICATING
I. Congenital lesions	**I. Congenital lesions**
A. Aqueductal stenosis	A. Arnold-Chiari
B. Atresia of the foramina of Magendie and Luschka	B. Encephalocele
C. Masses	C. Leptomeningeal inflammation
1. cysts	D. Lissencephaly
2. tumours	E. Congenital absence of arachnoid granulations
3. malformations	
II. Acquired lesions	**II. Acquired lesions**
A. Aqueductal stenosis	A. Leptomeningeal inflammation
B. Ventricular inflammation	1. infections
C. Masses	2. hemorrhage
1. tumours	B. Masses
2. non-neoplastic masses	1. tumours
	2. non-neoplastic masses

1.1.3 Treatment

Current treatment of hydrocephalus consists, almost exclusively, of surgical therapy. In a few cases, if the hydrocephalus is caused by a discrete obstructing mass, e.g., a tumour, the mass may be surgically removed and normal CSF flow may then be restored. In such cases, the hydrocephalus is usually temporary and easily relieved. However, in most cases, this is not possible and more definitive treatment of the

hydrocephalus itself is required. There are currently two main forms of surgical treatment for hydrocephalus and the principle behind each is to bypass an area of obstruction of CSF flow.

1.1.3.1 CSF Shunt

A CSF shunt is a silastic tube that serves to drain CSF from the ventricular system into some other part of the body where it can be absorbed. This represents, far and away, the most common form of treatment of hydrocephalus. In its most usual form, the CSF shunt consists of a silastic proximal (ventricular) end that is inserted through the skull (via a small burr hole), through the brain substance, and into the cerebral ventricles. This is then connected to a valve mechanism on the skull surface. The valve mechanisms vary, but in general, they serve to control the amount of CSF that flows out of the ventricle by regulating the pressure. This valve is then connected to a second, much longer, distal silastic tubing that is tunneled just underneath the skin to drain into the peritoneal cavity of the abdomen (ventriculoperitoneal (VP) shunt). Less commonly, the distal tubing may drain into the right atrium of the heart (ventricuolatrial (VA) shunt) or the pleural cavity of the chest (ventriculopleural shunt). In any case, the idea is to drain the CSF from the ventricles into another body cavity that will then absorb the CSF. In a small number of cases, it may be possible to drain the CSF from the lumbar spine area into the peritoneal cavity (lumboperitoneal (LP) shunt). This is becoming a less and less common procedure because of its associated complication rate.[10, 11]

While the principle of a CSF shunt sounds quite simple in theory, it is far from a cure for hydrocephalus. Shunts are associated with a significant failure rate that requires re-operation. While most such failures occur in the first 1 or 2 years, children are at life-long risk for developing shunt failure and needing further surgery.[12-14]

1.1.3.2 Endoscopic third ventriculostomy

Endoscopic third ventriculostomy (ETV) is a relatively new technique. This technique is feasible for only a very select group of children with hydrocephalus. Specifically, it is usually reserved for non-communicating hydrocephalus in which the obstruction occurs at or distal to the level of the cerebral aqueduct. The procedure involves placing a small endoscopic camera through a burr hole in the skull, through the brain substance, and into the cerebral ventricles. Once inside the ventricular system, the camera and probe are carefully manipulated through the foramen of Monro into the third ventricle. A small hole is then created in the floor of the third ventricle. This hole serves as the bypass outlet to let CSF flow around the obstructing lesion and be reabsorbed by the normal surrounding brain mechanisms.

This procedure is technically much more difficult than a CSF shunt. As such, it does carry a small risk of significant peri-operative morbidity, including severe brain hemorrhage. Even if it is performed successfully, a number of the patients eventually develop failure of the third ventriculostomy and require re-operation (either a repeat attempt at third ventriculostomy or insertion of a CSF shunt). A further concern

regarding third ventriculostomy is that, even in long-term successes, the ventricular system rarely returns to a normal size. There has been an on-going debate as to whether this continued ventriculomegaly will have a negative impact on the child's cognitive development. Although most earlier studies support this contention, the issue has not yet been resolved, especially with regard to ETV.[15-19]

1.1.4 Complications of Hydrocephalus

Hydrocephalus is associated with potentially significant complications and far reaching disability. These not only affect the developing child, but, in many cases, extend into the patient's adult life. There are a number of areas of dysfunction that need to be considered.

1.1.4.1 Mortality

Although the mortality in the treatment of hydrocephalus has been greatly reduced in recent years, recent series report mortality rates in the 5-14% range.[20-22] Many of these deaths occur in a delayed fashion, since peri-operative mortality has been substantially reduced to under 2%.[20] Many of these deaths are due to shunt infection or shunt failure.[21-23]

1.1.4.2 Shunt Infection and Malfunction

Shunt malfunction can result from shunt obstruction, shunt disconnection, shunt over-drainage, or shunt infection. Approximately 40% of patients will develop shunt malfunction in the first year after shunt insertion.[24] These usually require hospital admission and surgical revision. It is estimated that the mean duration of hospital stay associated with such conditions range from 12 to 22 days, depending on the etiology.[4] Overall, approximately 80% of shunted children will suffer at least one shunt malfunction during their life.

Shunt infection is a particularly feared complication that accounts for roughly 15-19% of shunt failures.[20, 22] The majority occur in the first year following shunt insertion.[25, 26] Shunt infection is a major cause of mortality in hydrocephalus, accounting for between 10-50% of deaths, depending on the series.[21, 22]

1.1.4.3 Cognitive Dysfunction

Various studies have reported on the intelligent quotient (IQ) of children with hydrocephalus. Hoppe-Hirsch et al. report that only 32% have IQ's above 90 while 40% have IQ's below 70.[21] This retrospective series found some association between the etiology of hydrocephalus and IQ: children with myelomeningocele fared the best while children who had post-infectious neonatal hydrocephalus did the worst. Donders et al. also found that the presence of neonatal infection, or other severe neonatal insults,

corresponded to a lower intelligence outcome.[27] A study of infantile hydrocephalus found that 36% of children were mentally retarded.[28]

School performance is, of course, also severely hindered. In a French study, although 60% were attending a normal school setting, over half were at least 1 to 2 years behind their peers. Thirty-one percent were attending special schools and 9% were considered ineducable.[21] A British study found similar results with 53% of their patients attending normal schools and children with post-infectious hydrocephalus performing the worst.[20]

A great deal of detailed work has been done exploring the language deficits experienced by children with early onset hydrocephalus. In a selected group of children with hydrocephalus between 6 and 15 years of age, and all with IQ's over 90, a significant deficit of reading comprehension has been described compared to controls.[29] This group also demonstrated impairment of oral comprehension and narrative discourse.[30-32]

1.1.4.4 Physical Dysfunction

The level of physical disability is sometimes difficult to quantify given the nature of the outcome measures used in the hydrocephalus literature. Hoppe-Hirsch et al. reported that, at long-term follow-up, 60% of patients had evidence of neurological motor deficit on physical exam.[21] Hydrocephalic children have demonstrated lower scores on measures of fine motor, visual-motor, and spatial skills compared to controls.[33, 34] Furthermore, 11-25% of children with hydrocephalus are left with some visual or hearing deficit.[21, 28]

1.1.4.5 Epilepsy

The incidence of seizures in patients with hydrocephalus varies, but has been reported at 30-48%, with higher rates in children with post-infectious hydrocephalus.[28, 35-37] Of all hydrocephalic children with seizures, 71% developed seizures following the initial treatment of their hydrocephalus, with most beginning very early on in the post-operative course. Of patients with seizures, between 40-86% have recurrent, poorly controlled epilepsy despite medication.[21, 36] These seizures occurred rather frequently with many having multiple seizures every week. Patients whose seizures began following treatment of hydrocephalus appeared to have more severe epilepsy than those whose seizures began before treatment. Half of all patients with seizures required at least one hospital admission for treatment of a severe, refractory seizure episode. A greater incidence of epilepsy was noted in children who had had more shunt complications. As well, it appears that seizures are associated with a very bad prognosis, with over half these children being institutionalized.

1.1.4.6 Psychological Impairment

It is recognized that many hydrocephalic children experience psychological problems. One study describes psychological problems in 80% of children, one-third described as "severe", with no further description of their terminology.[21] Donders et al.

found that in a minority of hydrocephalic children (16%), there was evidence of specific adjustment problems, including oppositional and acting-out behaviour and attention-deficit hyperactivity disorder.[38] Others have described markedly more behavioural problems, especially associated with hyperactivity, in children with hydrocephalus compared to controls.[39, 40] Children with hydrocephalus also appear to have lower perceptions of self-competence.[40]

1.1.4.7 Headache

Headache is a relatively frequent complaint of patients with hydrocephalus, being self-reported as the most common perceived side-effect of shunts.(E. Fudge 1999 – Hydrocephalus Association survey, unpublished data) Stellman-Ward et al. found the incidence of chronic headache in hydrocephalic children to be 58%, of which 22% were consistent with migrainous headache.[41, 42]

1.1.4.8 General Concerns

The Hydrocephalus Association recently conducted an informal survey of its members (E. Fudge, 1999 – unpublished data). While the methodology does not reach the high standards of many scientifically conducted surveys, the results do provide a very enlightening introduction to some of the concerns of people with hydrocephalus. In total, 422 people responded (63% response rate). Most of the patients (73%) were under 20 years of age and many required proxy responses from their parents. Nearly half the patients (49%) had between 1 to 5 shunt revisions and 18% required more than 5 shunt revisions (3% had had more than 20). The respondents expressed great concern over the shunt. The greatest source of concern was the need for further shunt revision, but also included concerns such as not knowing if the shunt is working properly, possible brain injury, the effect on their lifestyle, long-term complications, and fears about pregnancy. Headache was reported as the most common side-effect they associated with the shunt (31%), followed by problems with balance (19%).

1.1.5 Summary

Hydrocephalus is a common, debilitating condition that affects children of all ages. The disease has the potential to affect virtually all facets of a child's life. As the most common single condition seen by most pediatric neurosurgeons, there is great interest in learning more about the condition. This requires, of course, proper measures of clinical outcome. The following section will detail the current methods being used by researchers to measure outcome in clinical studies of pediatric hydrocephalus. Based on this, a case will be made for the need for a better outcome measure for this group of children.

1.2 Currently Used Hydrocephalus Outcome Measures

1.2.1 Introduction

Hydrocephalus has long been a focus of research within neurosurgery and in other specialties, as well. However, the neurosurgical research has mostly involved dealing with purely surgical outcomes, such as shunt infection or shunt failure. The number of studies to have included other outcomes, particularly patient- or family-based outcomes, is exceedingly rare. A number of other medical specialists, especially pediatric neuropsychologists, have taken an interest in the long-term effects of hydrocephalus. Much of their research has documented, using a battery of specific neurodevelopmental tests, the degree of deficit experienced by many children with hydrocephalus. While these tests can document specific deficits in great detail, they do not provide an overall picture of the disability nor do they provide any significant input from the family. In order to thoroughly document the current usage of outcome measures in pediatric hydrocephalus, a review of all relevant literature published from 1991 to 1999 was performed. This period was chosen since it was only relatively recently that the use of health-related quality of life measures or other health status measures has gained wide acceptance, particularly in the field of neurosurgery. The details of this literature review have been published elsewhere.[1] Of interest, the reader is alerted to a few key articles from before this time period: Dennis et al. 1981, Donders et al. 1990, Query et al. 1990, and Renier et al. 1988.[17, 27, 43, 44] These earlier articles used outcomes very similar to the more recent sample.

Of the 36 studies that met the inclusion criteria and reported on *non-surgical outcomes*, 17 (47%) used some reporting of symptoms or signs from a neurological examination as an outcome. Nine (25%) used some indicator of school performance as a reported outcome. Three (8%) reported results of vision and hearing tests. The majority of studies (28, 78%) used some form of neuropsychological testing. Overall, 43 different types of neuropsychological tests were used a total of 112 times in these 28 studies, for a mean of 4 tests per study.

1.2.2 Deficiencies of Currently Used Outcome Measures

There were a number of striking findings that highlight some of the deficiencies in the currently used outcomes measures which warrant further discussion.

1.2.2.1 Lack of use of non-surgical outcomes in the neurosurgery literature

There was a noticeable paucity of studies published in the standard neurosurgery journals. Of the 36 studies, only 12 were published in journals aimed primarily at neurosurgeons (Acta Neurochirurgica, Child's Nervous System, Journal of Neurosurgery, Neurosurgery, and Pediatric Neurosurgery). Neurosurgeons are far and away the medical

specialists most associated with the treatment of hydrocephalus and it is neurosurgeons who, aside from providing acute care, provide the bulk of long-term follow-up care. For many patients with hydrocephalus, neurosurgical follow-up continues throughout their life. Admittedly, the focus of most neurosurgical research has been acute care and the minimization of surgical shunt complications. All articles related to those areas were excluded from the above analysis. However, it is still rather surprising to see that most of the literature documenting the non-surgical outcome of children with hydrocephalus is not found in the journals most likely to be read by neurosurgeons.

It is difficult to interpret this observation. Given the long-term relationship that most pediatric neurosurgeons develop with their hydrocephalic patients, it is not likely that this lack of research spawns from lack of academic interest. One possible explanation is the complexity involved in using currently available non-surgical outcomes, as witnessed by the 43 different types of neuropsychological tests used by other authors. Most neurosurgeons lack familiarity with the implementation and interpretation of such testing.

This possible lack of acceptance of such outcome measures by neurosurgeons is an important issue. One of the main goals of most clinical research is, ultimately, acceptance of the findings by the clinical community. In the absence of this, the results of any research will have no impact and no possible benefit to patients. This must be recognized and any new measure that is proposed should be acceptable and credible to neurosurgeons.

1.2.2.2 Overwhelming number of neuropsychological tests

A further finding from the above study was the vast number of different tests used to document non-surgical outcomes in hydrocephalus. Aside from simple descriptions of symptoms and signs, school performance, or vision/hearing tests, a staggering total of 43 different types of neuropsychological tests were used a total of 112 times. Among the studies that used these tests, each child was subject to a mean of 4 different tests per study. These tests are well established amongst neuropsychologists and many have been used for decades. However, their use does have many drawbacks. Many of these tests are rather involved and demand a fairly substantial time commitment. As well, the implementation and, very importantly, the interpretation of such tests require neuropsychologists with special training in dealing with children. This is but one limitation of this form of testing. A further limitation is that while each test is perhaps very good at detailing a very specific spectrum of deficit, they fail to provide an overall sense of the child's health status. This is not to say that these tests do not serve a useful purpose. If one is interested in specifically exploring language deficits in children with hydrocephalus, for example, there are many well-established tests that may be used. However, from the neurosurgeon's clinical perspective, there is likely more interest in determining the child's overall health status, taking into account the numerous different domains that this encompasses. These neuropsychological tests do not allow for the creation of an overall picture of the child's health status, even when performed as a serial

battery of tests.

1.2.2.3 Lack of disease specificity

The outcome measures used lack disease specificity. There are issues that are specifically problematic to patients with hydrocephalus that are not addressed by these measures. This is really a question of content validity and face validity. This brings into issue acceptance of the outcome measures, not only by neurosurgeons, but by patients and their families, as well.

1.3 New Outcome Measurement

The previous discussion has established that the currently used outcome measures in pediatric hydrocephalus are deficient as measures of disease-specific health status. What then are the desired properties of a proposed new instrument? A useful starting point is to consider what the potential uses of the instrument would be.

1.3.1 Uses for a New Outcome Measure

The development of a new outcome measure for pediatric hydrocephalus will have widespread application for use in any future studies to examine hydrocephalus treatment or general outcome. It will thereby provide a standard, validated measure that can be used to compare the long-term results of different forms of shunting devices or endoscopic techniques, for example. Specific examples of uses include:

1.3.1.1 Outcome in prospective studies

In a recent randomized controlled trial comparing 3 different shunt valves all recorded outcomes were purely surgical, e.g., shunt failure.[24] A disease-specific health status measure would have been an ideal secondary outcome in such a study. Although there were no differences noted in the rate of shunt failure among the three groups, no other assessment was done to see if there were differences in other aspects of the child's health status.

Another specific use would be as a primary outcome measure to compare results after third ventriculostomy versus routine CSF shunting. This is an area of great controversy and the advantages of ETV have not been well explored in a direct comparison to shunts. Although the surgical outcomes would be of great interest, it would also be extremely useful to have some other measure of the child's well being. This is especially relevant in this situation since many proponents of ETV argue that it eliminates many of the concerns patients and families have with shunts, such as fear of shunt failure. However, there are those who argue that ETV does not sufficiently reduce ventricular size and may, in fact, lead to greater overall cognitive impairment. A disease-specific health status outcome may be able to shed light on some of these issues.

1.3.1.2 Study long-term outcome of children with hydrocephalus

As was documented in previous sections, we have little information regarding the overall health status outcome of children with hydrocephalus. While there has been research into specific areas of cognitive or behavioural deficit, none have been able to provide a comprehensive picture from the families' perspective. With a new outcome measure, it would be possible to study the long-term outcome of a large cohort of children with hydrocephalus. Not only would this provide the more comprehensive picture of outcome that is so desirable, but it would also allow us to make the first steps in establishing the early prognostic factors associated with improved health status. Ultimately, such work would allow surgeons to be better able to prognosticate about a child's health outcome at the very early stages of the child's illness. As well, since the outcome is from the families' perspective, surgeons may be better able to explain the possible health status in terms that are particularly meaningful to the parents. This instrument would expand our current thinking of treatment success beyond simply the incidence of shunt malfunction, shunt infection, or other objective clinical measures.

Another specific area of use would be to study the long-term outcome of children with fetal hydrocephalus. This is an area that was of great interest in the 1980's with the enthusiasm for performing fetal surgery to correct *in utero* hydrocephalus. However, the initially poor results led to a moratorium on such work. Recently, with advances in technology, a new wave of interest has spread in fetal surgery. With this comes renewed interest in assessing the outcome of children with fetal hydrocephalus.

1.3.2 Properties of a New Outcome Measure

Given what has been discussed about currently used and available outcome measures, it is useful to detail the desired properties of a proposed new outcome measure for hydrocephalus.

1.3.2.1 Scientific Requirements

1.3.2.1.1 Reliability and Validity

The new measure must, at a minimum, be a reliable instrument that is able to consistently discriminate between subjects and a valid instrument that actually does measure the child's health status. These psychometric properties will be discussed in greater detail in later sections.

1.3.2.1.2 Need to discriminate between groups with similar disease burdens.

Many of the previously established outcome measures are able to discriminate between chronically ill and well children and some are able to discriminate within the group of chronically ill children. However, within the group of children with hydrocephalus the issues that discriminate may be different. For example, while differences in major areas such as cognitive functioning or level of self-care may be

detected by other outcome measures, they ignore other issues such as seizures, chronic headaches, and concerns relating to the shunt. These issues may be better able to finely discriminate between those children who otherwise appear to be similar. This is one of the main arguments for a disease-specific instrument. Since the main use of the new measure will be in conducting research (rather than in individual patient management), the instrument must be able to discriminate between groups of patients, rather than individual patients.[45] These desired properties are in contrast to an evaluative instrument, for which the main interest lies in distinguishing changes within a group or patient over time.[46] It is expected that this instrument will be used as an outcome in prospective trials of competing interventions. As was discussed earlier, hydrocephalus frequently presents in an acute or sub-acute fashion. For example, in many cases it presents in a critically ill premature neonate who develops hydrocephalus while still in the neonatal intensive care unit. It may also present acutely in a child who was otherwise healthy until a rapid deterioration 1-2 weeks prior to diagnosis. Both of these children would receive urgent surgical therapy at diagnosis. The interventions one performs for hydrocephalus, unlike in asthma for example, are applied in relatively urgent situations to an acute deterioration, rather than to a stable, chronic disease process. It is easy to see how it would be extremely difficult and meaningless to attempt to obtain some measure of baseline health status prior to intervention. Although there are cases in which children will experience a more gradual decline in level of functioning, this represents the minority. Even in these cases, the history is one of deterioration, without, usually, a period of stable baseline dysfunction from which one could obtain a meaningful health status measure. The point is that without useful baseline measures, one can not look for a *change* in instrument scores after the intervention. Rather, in the vast majority of cases, only post-intervention scores will be available for comparison. Therefore, the instrument that is needed is one that has good discriminative properties - one that is designed to measure *between-subject differences* on an underlying continuum of health status.

The instrument will be used for group-level, rather than individual-level, measurement. The reason for this is that its primary purpose is to be used as a research tool, studying patient cohorts. In general, individual-level measurements would be important in situations where the results would affect clinical decision-making. However, in the majority of cases of long-standing hydrocephalus, clinical decisions are usually acute surgical decisions made on the basis of rather dramatic physical signs and symptoms, e.g., severe headache, papilledema, deterioration of vision, deterioration of consciousness, or changes in the brain imaging. Therefore, the uses of a health status outcome measure for individual patient management in hydrocephalus would be limited and this will not be the intent of this new instrument.

1.3.2.1.3 Comprehensive coverage

It is important that the outcome measure cover all components of health status that are considered important to neurosurgeons and patients.[47] As already mentioned, previous measures have ignored some areas that may be particularly important to the

population of children with hydrocephalus.

1.3.2.2 Practical Requirements

1.3.2.2.1 Easy to use and score

It is very important that the new outcome measure be easy to use and score. Ideally, it should be short enough to be completed in less than half an hour and should be self-administered. This will increase its use by investigators and likely increase patient and family compliance. If one develops an outcome that is not used, then it serves no purpose.

1.3.2.2.2 Well accepted by neurosurgeons

The new outcome measure should be aimed at the needs of neurosurgeons. Specifically, it should be well accepted by them as scientifically credible and important.[47] Once again, this is important if the instrument is to be used widely.

1.3.2.2.3 Well accepted by patients & families

It is also important that patients and family view this new outcome measure as credible. This speaks to the face validity of the measure.[47]

1.3.2.2.4 Yield an overall score

It is necessary for the new measure to provide a single overall score for health status. This would be advantageous for its use as an outcome in comparative clinical studies and it would then be feasible to attempt to provide some clinical interpretability to certain ranges or differences in score (in terms of their clinical magnitude and relevance).

2. Theoretical Background

This chapter will provide a review of some of the theoretical concepts that are relevant to the development and testing of a new health status outcome measure for hydrocephalus. The practical methods involved will be discussed in later chapters. Some theoretical issues will not be discussed here (e.g., the use of proxy respondents and theoretical aspects of validity testing) because the author has already dealt with these in previous work.[1]

2.1 Measurement Theory

The development of a health status outcome measure is based on certain implicit assumptions about the characteristics of what is being measured. However, it is useful to explore these assumptions in an explicit manner, especially in the context of the current outcome measure. The set of assumptions that follow are those that comprise *classical true-score theory (CTST)*.[48]

2.1.1 Assumptions of Classical True-Score Theory

As stated by Allen and Yen, there are seven assumptions of CTST.[48] However, for the hydrocephalus health status outcome measure, there are five that are most relevant and will be discussed.

2.1.1.1 Assumption One: $X = T + E$

In this assumption, X is the observed score from the health status measure for a respondent, T is the true score for that respondent, and E is the error of measurement. This is the assumption of additivity. That is, the observed score is the result of summation of the true score and the error, rather than, for instance, multiplication. The error of measurement represents the effect of random error on the observed score. Sources of such error include poorly standardized instructions, differences in testing environment (e.g., home versus hospital), or fluctuations in the respondent (e.g., changes in mood or motivation).[49]

2.1.1.2 Assumption Two: $\mathcal{E}(X) = T$

This assumption states that the expected value, or population mean, of an infinite series of independent observations on a respondent with the same measure is equivalent to the true score for that respondent. In other words, the true score will always remain a theoretical construct only. A corollary to this assumption is that the measurement error associated with each observation must be entirely random, so that over a series of infinite

observations it has no net effect on the observed score, i.e., the observed score mean approximates the true score.

2.1.1.3 Assumption Three: $\rho_{ET} = 0$

Under this assumption, the measurement errors and true scores for a population of respondents are uncorrelated.

2.1.1.4 Assumption Four: $\rho_{E1\text{-}E2} = 0$

This states that the measurement error for a respondent taking test 1 (E1) is not correlated to the measurement error of a separate test 2 (E2).

2.1.1.5 Assumption Five: $\rho_{E1\text{-}T2} = 0$

Under this assumption, the measurement error on test 1 (E1) is uncorrelated with the true scores on test 2 (T2).

The previous assumptions form the basis for many of the psychometric properties of the instrument that will be measured, especially reliability. However, there is another measurement theory that is also relevant to the development of a new outcome measure, generalizability theory. This will be discussed in a later section (see **Section 2.2.3**).

2.2 Reliability

2.2.1 Interpreting the Reliability Coefficient

Reliability can be interpreted in different ways. In particular, the reliability coefficient has several different theoretical interpretations that stem from CTST.[48] Allen and Yen describe several of these interpretations of the reliability coefficient (ρ). These include the following:

i. The coefficient ρ can be interpreted as the correlation between observed scores using the measure and observed scores using a parallel test form. A parallel test form is an alternative measure that provides the same true test score and error variance as the measure under consideration. This is largely a theoretical consideration, since the true score for any test is itself a theoretical construct (see above discussion of CTST) and can not be measured directly.

Related to this interpretation of ρ is that it represents the squared correlation between a measure and a parallel test form. That is, it is the "proportion of variance in one test score explained by its linear relationship with scores on a parallel test".[48]

ii. The coefficient ρ can also be interpreted as the ratio of the true-score variance to the observed-score variance. That is, the proportion of observed variance that is due to

true-score variance. If a measure were perfectly reliable, all observed variance would be due to variance in the true score and $\rho = 1$. In this situation, there would be no error associated with an observed score.

iii. The coefficient ρ can be interpreted as the square of the correlation between the observed scores and the true scores. In the absence of error, this correlation would be one. This correlation also represents the upper limit of the measure's correlation with any other measure. That is, a measure can not correlate more highly with any other measure than it can with its own true score. This interpretation, therefore, has implications beyond simply reliability, since it represents the upper limit of correlations that might be expected in criterion validity testing with another measure.

Related to this, the coefficient ρ can also be interpreted as 1 minus the squared correlation between the observed scores and the error scores. Once again, in the absence of error, the correlation would be zero and the reliability would be perfect.

iv. Alternatively, the coefficient ρ can be interpreted as one minus the ratio of error-score variance to observed-score variance (or the proportion of observed-score variance attributable to error-score variance). This interpretation provides some practical insight into reliability estimation. That is, the reliability estimate can be substantially influenced by the degree of variance in the test subjects. Because it is in the denominator, a test sample with greater heterogeneity and variance, in the presence of roughly equal error variance, will produce a greater estimate of reliability than a more homogeneous, restrictive sample.

2.2.2 Estimating the Reliability Coefficient

Given the many different theoretical interpretations of the reliability coefficient, it is not surprising that there are different methods of estimating the coefficient. The three most common methods for estimating reliability are test-retest, parallel forms, and internal consistency. Each is ideally suited for different situations and, in general, they provide somewhat different estimates.

2.2.2.1 Test-Retest Reliability

This method involves comparing the results of a test given to the same respondent on two separate occasions. When the correlation of these scores are calculated, this, in effect, exploits the interpretation of coefficient ρ as a correlation between observed scores on parallel tests. Strictly speaking, these tests are not actually parallel, but, in fact, identical except for their temporal spacing. There are, however, other ways in which the scores can be compared. Most commonly, for example, one might use an intra-class correlation coefficient (ICC). This method partitions out the observed variance to determine how much is attributable to true variance. This exploits another definition of

the reliability coefficient ρ: the ratio of the true-score variance to the observed-score variance.

The main difficulty with the test-retest method is the potential for carry-over effect.[48] That is, the first testing may influence the second. This could occur if respondents remember the questions from the first test, if they seek advice in the interval between testings, or if they change their attitude about the testing. The degree to which these occur affects how independent the observations are and how meaningful a comparison will be. A further problem is that if the trait one is measuring changes quickly relative to the time interval between tests, then the expected true score on the two tests will be different even in the complete absence of carry-over effects. In this situation, the tests are no longer parallel tests.

2.2.2.2 Parallel-Forms Reliability

In parallel forms reliability, the correlation between two parallel tests is used as the estimate of reliability.[48] The problem with this, however, is that it is difficult to show that two tests are parallel because this relies on a theoretical construct (namely equality of the true scores of both tests). In actuality, alternate test forms are used. These are tests that have been constructed to make them essentially parallel with roughly equal mean scores and variances.

One of the problems with alternate test form reliability is also carry-over effect, which is a concern because the alternate test forms are so similar. Therefore, just as with test-retest reliability, the interval between testing is important.

2.2.2.3 Internal Consistency Reliability

Internal consistency reliability estimates are based on a one-time administration in which a single test is essentially broken down into two parts. This is done in such a way that the two parts are as close to each other as possible. In this manner, the two halves are either parallel test forms (i.e., equal true score and equal error variances) or, more commonly, they are tau (τ) equivalents (true scores are equal with an additive constant). These definitions are provided as part of CTST. The reliability coefficients for the entire test are provided by either the Spearman-Brown formula (for parallel tests) or coefficient alpha (for τ-equivalents). In the more general case, a test can be divided into more than halves (perhaps into thirds, eighths, etc.). In this situation, if each of the components are τ-equivalents, then an alpha coefficient can be calculated. The most practical example is for each individual item in a measure to represent an individual component.

Although the obvious advantage of internal consistency reliability is the need for only a single administration, the main limitation is its lack of suitability for certain types of tests. Most obviously, highly time-dependent tests of speed are not appropriate candidates for internal consistency reliability

2.2.3 Generalizability Theory

When assessing the reliability of a measure, it is likely that there are different sources that might contribute to lack of reliability. These include, for example, variables such as the time of measurement, the location in which a measure is completed (e.g., home versus hospital), or the type of respondent (e.g., mother or father). The effect of each of these on reliability could be tested by conducting several different reliability studies. However, it would be more efficient to assess these within one larger study design. This would also allow one to test which, if any, of the different variables are contributing to a lack of reliability and to quantify the effect that this is having. By identifying the responsible variable(s) then steps can be taken to remedy the situation. This might involve a modification in the measurement itself (to reduce the effect of the variable) or it might involve changing the way in which the measurement is performed (e.g., limiting it to certain types of responders or certain locations). This mode of studying the effect of multiple variables on the reliability of a measure is called a generalizability study (G study).[48, 50] This is based on generalizability theory (G-theory), which is first credited to Cronbach et al. in 1963.[48, 50-52] In classical true-score theory (see **Section 2.1**), it is the assessment of reliability as an indicator of how well observed scores represent true scores that is of greatest interest. However, in G-theory "attention focuses on dependability, or how well observed scores allow generalization about behaviour in a defined universe of situations."[53] As stated by Cronbach, in the seminal 1963 paper entitled *Theory of generalizability: A liberalization of reliability theory*, "an investigator asks about the precision or reliability of a measure because he (or she) wishes to generalize from the observations to which it belongs".[51] The group of observations referred to by Cronbach actually defines the universe of observations that could produce a usable score within the context of the research. This universe, therefore, is variable and can be determined by the researcher to be as broad or constrained as reasonable. It depends on the potentially different circumstances which could produce a usable score. These circumstances may vary based on the observer, the time of the observation, or the place of the observation, for example. Each unique combination of circumstances might produce a somewhat different score and it is the average of the scores of each unique set of circumstances which yields the *universe score*. The universe score is the G-theory equivalent of the CTST true score.[53]

The design of a generalizability study first involves identifying the potentially relevant variables that might impact on the reliability of a measure. These are the within-subject sources of error, where "subject" refers to the object of measurement. These are, obviously, very dependent on the measure itself. Once the variables are identified, then the levels within each variable must be defined. For example, for a health status measure of children, a potential variable might be the parent who is responding to the measure. Therefore, the levels within that variable would be "mother" and "father". The design must then incorporate repeated completion of the measure under every unique combination of the various levels of each identified variable. For example, if there are 2

variables with 2 levels each, then each measure must be repeated 4 times. This is an attempt to sample the universe of potential scores.

The statistical analysis of G studies can be broadly divided into two related goals: assessment of specific reliability coefficients and exploration of sources of within-subject error.[50] The latter involves a repeated-measure analysis of variance (ANOVA) which attempts to partition variance according to the within-subject variables. Those that are identified as "significant" sources of error represent variables which appear to be contributing to poor reliability. These variables need to then be accounted for in the use of the measurement, either by changing the measure itself or restricting the use of the measure to nullify the effect of these variables. The degree to which the truly important sources of error have been considered in the study design is reflected by the magnitude of the residual error: the smaller the error, the more that has been accounted for by the variables under consideration.

Individual reliability coefficients can then be determined to reflect unique measuring situations. These are called coefficients of generalizability and each measure will have numerous such coefficients, depending on the number of variables being considered. For example, a reliability coefficient can be calculated to reflect different observers making measurements at the same time (standard inter-rater reliability) or the same observer making repeated measurements on different days (standard intra-rater reliability) or, even, some combination of these. These types of reliability tests are simply restricted cases of the more general situation which was part of the G study.[50]

An extension of a G study can also allow for estimating the effect a change in the measurement protocol would have to the reliability coefficient.[50] For example, an estimated reliability coefficient can be calculated for the average measurement of three observers performed on different days (as opposed to just a single observer). Various different changes can be tested to see which, if any, might lead to acceptable levels of reliability.

Within G theory, two types of error variances are recognized: absolute and relative error variances. Absolute error variance is defined as the variance of the difference between a subject's observed and universe scores.[54] It can be estimated by summing up all other variance elements other than that due to the subject itself. Absolute error is useful when the measurement results of each subject will be used independently, rather than in comparison with other subjects in the group of study.[55] Relative error variance is defined as the variance of the difference between observed and universe scores *relative to the population means* for observed and universe scores.[54] It can be estimated by summing all the variance elements attributable to interactions with the subject. This includes some, but not all, of the variance elements that comprise absolute error. Therefore, relative error variance is smaller in magnitude. Relative error is useful when the measurement of each subject is to be ranked and compared to the entire sample.[55] This would be the case, for example, in measuring the performance of a class of students. In most clinical medical settings, absolute error is more relevant. The chosen error variances can be compared to the subject variances to provide some estimate of their

relative magnitude. Most commonly a signal-noise ratio is calculated by dividing the subject variance by the error variance.[54] The higher the signal-noise ratio, the greater is the ability of the measurement scenario to make the intended discriminations among subjects.

2.3 Item Scaling[§]

In developing a new instrument to measure health status in hydrocephalus, decisions will need to be made about the way in which the items are scaled. There are various options regarding item scaling. This requires consideration of both how the item is to be worded and what the response options will be. Integral to this decision, also, is how to create a final score.

2.3.1 Time Specifications

When asking questions about health status it is important to define a clear time reference upon which the respondents will base their answers. For example, Juniper et al. asked respondents to answer HRQL questions based on the previous 2 weeks.[56] For this new instrument, all questions will be asked with reference to the previous 4 weeks. It is felt that the concept being measured should be relatively stable, barring any significant interceding events. This period should allow for an adequate representation of the child's daily functioning in most circumstances but is short enough that issues such as developmental changes need not be of concern. The 4 week time reference has also been used successfully by others.[57, 58]

2.3.2 Wording of Item

One of the simplest ways to word an item is to frame it as an Agree/Disagree statement. For example, the item may be as simple "Your child is able to make new friends". The corresponding response would then be dichotomous. Although simple, it ignores the greater complexities associated with issues of well being and health-status. It is likely that parents would prefer to be able to answer items on some continuum, since many issues are not simply either present or absent.

Previous scales that have assessed quality of life in children have formatted the items such that they were phrased in the form, "To what extent do you feel that you .....e.g. are able to make your own decisions?". This was used successfully in the

---

[§] This section is taken largely from the candidate's Masters thesis[1] and is not being presented here as original new work. It is provided simply to allow the Doctoral thesis to be used as an independent document.

instrument designed for children with spina bifida.[59] This then allows the respondent to answer with more variety than simply Yes or No. This is advantageous for at least two reasons. Firstly, this acknowledges that when answering about issues that relate to well being and health status, many of the issues can not fairly be answered in a dichotomous fashion. Rather, many represent traits for which there truly exists a continuum. The second advantage is based again on the measurement properties of the instrument. By allowing for a continuum of responses, this has the effect of spreading out the data. This may potentially increase between-subject variability. With all else being equal, this will tend to increase the reliability of the instrument.

A further consideration is to keep the wording of the items very simple and clear. Even though the instrument is intended for use by adults, there may be varying degrees of educational background and facility with the English language.

2.3.3 Response Options

Given the wording of the items, the response options will need to allow the respondent to express a continuum of option. This will range from a very positive response at one extreme to a very negative response at the other extreme. Perhaps the most common way is to use a modified Likert scale. This allows for responses on a continuum that ranges from "Not at all" to "To a very great extent", for example. These anchor phrases attempt to show the extremes of opinions. In between these anchors, one has the option of breaking the scale up into discrete options or leaving this up to the respondent to place a mark on a straight line (visual analogue scale - VAS), for example. The problem with the latter is that it becomes a more tedious task to quantify the response to each item. Presumably this would require an actual measurement for every response. It has been suggested that the perceived precision of the VAS is false, since it can be argued that the human mind artificially breaks up the line into discrete sections anyway. The number of discrete sections is not clear, but classical work by Miller has suggested that people have, in general, the ability to discriminate up to about seven levels.[60] Another problem with the VAS is that subjects may find it more difficult to fully grasp the concept of the VAS and that a longer training period may be needed to be able to use it competently.[61]

Given these disadvantages for VAS, a scale with discrete options seems preferable. The number of response options chosen should again take into consideration theoretical and practical considerations. In theory, increasing the number of response categories can potentially increase the reliability coefficient. There is a theoretical limit to this, such that having more than between 7 and 10 categories yields little extra benefit.[62]

Looking at the overall situation, the total tradeoff is between instrument complexity and instrument simplicity. The basic factors that decide this are the response options and the number of items. Increasing either will increase instrument complexity, but will also increase the theoretical reliability. A tradeoff can be made by increasing item numbers, while trying to limit the number of response options, for example. This

will allow a greater sampling of content and increase the internal consistency. In some situations, the area of sampling may be restricted so that the only option is to increase response options. This should not be a limiting factor for measurements of health status, a concept that has a very vast domain of sampling.

Another issue is whether to provide descriptions at each numbered option or, simply, just the extreme options. Hughes and Dodder found that when measuring alcohol consumption the presence and type of individual descriptors did seem to alter the mean values of the responses.[63] However, the items were strongly correlated, regardless of the type of response option.

Taking all factors into account, for this new instrument, the response options will be placed on a 5-point ordinal scale anchored by the terms "Not at all true" at 1 to "Very true" at 5. Further descriptors will be used to anchor each response option in between these extremes, as well.

2.3.4 Consistency of Items & Responses

Most previous questionnaires have worded items in a consistent way, such that a strongly positive response means roughly the same thing for each item. This is an important issue, since a further option may be to intentionally reverse the wording or responses for certain items. This can serve as a double check of accuracy. For example, by repeating an item later on in the questionnaire, but this time reversing its wording or response options, one can test if the respondent was answering consistently, i.e., their answer was also appropriately reversed. If not, the reasons for this should be sought. Possible reasons include, lack of attention to the items or legitimate lack of understanding of the items. If this becomes a recurring event, then one must acknowledge that it is the fault of the instrument, which needs revision.

However, the potential drawback to this is perhaps more subtle. It may be viewed by respondents as a trick or a show of bad faith on the part of the investigators. The effect of this on their responses is difficult to know. One potential effect is lack of compliance or low completion rate of the instrument. Furthermore, by reversing certain items, this may serve to legitimately confuse some respondents. In such a situation, many of their answers may not truly reflect their actual feelings.

For this instrument it is felt that all items will be worded in the way that sounds the most natural. That is, no specific attempt will be made to convert a naturally negative item into a positively worded item. This approach should provide the least confusion to the respondents and make the questionnaire as simple to use as possible. While it will slightly increase the complexity involved in scoring the instrument, this should not be too burdensome a task.

A potential theoretical drawback to instruments having both positively and negatively worded items is that there may be a discrepancy in responses based largely on this division in wording. That is, using factor analysis, others have shown that the type of wording (positive or negative) has a strong influence on responses and that like-worded items group together very strongly.[64] At times this grouping may be so strong as to

obscure other correlations within the instrument. Although this finding is open to interpretation, it is possible that this represents a differential psychological response to positive or negative worded questions.

2.3.5 Summation

Once a format of the items and responses has been established, it must then be determined how a final score for the instrument will be calculated. The simplest way is a simple summative score, in which the score for each item (from 1 to 5) is summed over all the items, with a higher score presumably representing "more" of the concept. In this system, the scoring of negative items will be reversed in this summation.

More complicated options do exist for calculating the final score. Firstly, one might consider weighting the items. One rationale for this is that it may seem theoretically plausible that a certain item or domain is more important to the overall concept and, therefore, its score should be more heavily represented. However, as this is a relatively new concept that is being measured, the only empirical evidence of the relative importance of items comes from the item reduction phase of this instrument development. In that phase, items of least general importance will be systematically excluded. Therefore, the remaining items will likely be within the same range of importance to the concept. The issue of weighting is complex and the only thing that is clear is that it adds a significant amount of complexity and time to calculating the final scores. The theoretical benefits are not clear and are not universally accepted.[50]

A second form of weighting that is more subtle is how to deal with the different domains. It is quite possible that the grouping of items within the domains leads to an unequal number of items. This is, in a sense, a form of weighting, favouring the domains with the greater number of items. One option would be to calculate a score for each domain and represent those domain scores equally amongst all domains. This form of weighting discounts the importance of items in the larger domains. However, this would all be based on theoretical conjecture as to how these domains should be weighted. This is a tenuous argument, since the domains themselves are artificial groupings created by the instrument developers. There is no clear evidence, as yet, that they have any credibility as true distinct entities. As well, the items have been pre-selected to be those that fall within the same range of importance according to the target population, and it is they who truly define the concept. Therefore, the instrument likely will not benefit from further adjustment in the form of weighting, either for the individual items or the domains. This then brings us back to the original, and simplest, solution, which is to simply add the scores for each individual item to yield the final total.

2.4 Interpretability

Most health status instruments provide some numerical score that attempts to quantify the underlying concept which they are purported to be measuring. However,

different instruments will likely provide scores on quite a different metric. As well, even within a given metric, it is not usually evident what a given score means. For example, in clinical terms, how good (or bad) is the health status of someone who scores "50" on a particular health measure? And how much worse is that person compared to someone who scores "75"? How would one interpret a mean difference of "25" between one group of patients who received a particular treatment and another group who received a placebo? Are these differences "real"? Are these differences clinically meaningful? These are the types of questions that have led to active interest in ascribing interpretability to health status measures. This would allow for a conversion of the numerical score, from whatever type of metric it originates, into clinically relevant terms. Guyatt has highlighted what he considers the three requirements for the interpretation of a target health measurement instrument:[65]

i. availability of an independent standard(s)
ii. standards that are interpretable, i.e., "we must have an intuitive sense of the value"
iii. standards that are moderately or highly correlated with the target instrument

As one might expect, even when these criteria are fulfilled, the transformation is not easy and there still exists debate not only on the procedural aspects of this transformation, but on its fundamental concept as well.

2.4.1 The Minimal Important Difference (MID)

2.4.1.1 Definition of the MID

A frequent use of health status instruments is in providing comparisons of health status scores to determine differences. These comparisons could be of many types: within patient comparisons over time, within group comparisons over time, or between group/patient comparisons at a single point in time. In any case, the initial comparison involves determining if a true difference exists. However, the definition of a *true difference* is not absolute. In one sense, it may be defined purely statistically. That is, with the application of the appropriate comparative statistical test, e.g., t-test, and an agreed upon level of alpha-error, i.e., the p-value, one can provide a rather simple threshold, beyond which a difference may be considered a *true difference*. This concept of a statistically significant difference is well-accepted and has been used for many years. However, the difficulty comes in translating this concept into clinical decision making. Just because a difference is statistically significant, should it mandate a change in clinical practice? One of the major limitations of this concept is that it is dependent on numerous factors aside from the existence of a true difference. For example, the determination of a statistically significant difference will depend on the heterogeneity within the comparison groups relative to the score metric (and, therefore, the standard deviation of the scores) and the sample size of the groups. In the most extreme theoretical example, *any*

difference between groups could be shown to be statistically significant if a large enough sample size is recruited. The concept of statistically significant differences reduces itself, therefore, to a rather artificial one.

To deal with the limitations of pure statistical significance, the notion of the minimal important difference (MID - sometimes also called the minimal clinically important difference) has been championed. A widely quoted definition, by Jaeschke et al., of the minimal important difference is the "smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive costs, a change in the patient's management".[66] This is not the only definition of the MID and, as with any attempt to simply define a difficult concept, is subject to controversy. A more complete discussion of these controversies will be presented in a later section (see **Section 2.4.2**).

2.4.1.2 Measurement of the MID

That several methods of determining the MID have been reported in the literature, is, in itself, a testament to the ambiguity of the concept. A recent and thorough literature review by Wells et al. found nine reported procedures.[67] A brief outline of these procedures follows:

2.4.1.2.1 Patient perception: comparison to a global rating.

This approach is aimed at determining differences within groups over time. It has been used for the Chronic Respiratory Questionnaire (CRQ) by Jaeshcke et al..[66] Patients were followed for a period of time and asked to rate their self-perceived changes in the three domains of the instrument on a 15-point scale (ranging from "Almost the same, hardly any worse/better at all" to "A very great deal worse/better"). These global ratings were then compared to the change in the CRQ score, with the assumption being that the MID would correlate with a global rating of change of "Somewhat better/worse" or less.

There are several criticisms of this approach, and the author has highlighted some of these in previous work.[1] This method relies on patient recall of their previous health status, which may be subject to bias, and then it requires their ability to translate it into seemingly very similar terms.[68, 69] For example, the difference in interpretation between "About the same" and "Almost the same, hardly any worse/better at all" is tenuous. Therefore, one can question the psychometric properties of the global rating scale itself: how reliable is it and why does it seem that there is an *a priori* assumption of its validity?

2.4.1.2.2 Patient perception: patient conversation.

This method involves having patients with similar disease conditions interact with each other over a period of time and then have the patients rate their health relative to another patient. This was also done for the CRQ and the patients' response options ranged from "About the same" to "Much better/worse" than their fellow patient.[69] The clinical setting was a respiratory rehabilitation program and, in that sense, was rather

unique and lent itself to this technique. This procedure differs from the previous one in that it gives some insight into between-patient differences rather than within-patient differences. As well, it is not subject to recall bias. However, it does require two patients to have a prolonged exposure to each other in order to allow them to get a good assessment of the other's health. It is also difficult to know if patients were able to get an accurate sense of the other's health, even in this setting. Given the problems that an experienced clinician might have in assessing this, the task must be that much more difficult for a patient.

2.4.1.2.3 Clinician perspective: consensus development (Delphi).

This is a completely different method that involves surveying clinicians about their opinions about the MID level for an outcome measure.[67] Several expert clinicians were provided with basic data about an outcome measure (including means and standard deviations) and asked to provide their anonymous estimate for a MID for a given hypothetical randomized trial.[70] Each clinician then received the estimates of the other clinicians and was asked if they wished to revise their first estimate. This was repeated a third time, with the goal being some convergence in the clinicians' estimates. This provides a clinician-based estimate of a group-difference MID within the context of a specified randomized trial.

2.4.1.2.4 Clinician perspective: patient scenario scoring.

In this method, clinicians were provided with a description of the "average" of a certain outcome measure.[67, 71] They were then asked to consider a hypothetical treatment known to improve the outcome and without contraindication. They were asked to decide how much the outcome measure would need to change before they would recommend the treatment. In a second scenario, clinicians were told the typical percentage incidence for an adverse outcome and asked again to decide how much that percentage would need to be reduced (either in absolute risk reduction or number-needed-to-treat) before they would recommend the treatment. The scenario chosen was the use of propranolol to slow aortic aneurysm growth and they found that the MID varied substantially according to physician specialty and experience. As with the previous method, this is geared towards determining MID in the context of group comparisons, such as randomized trials, based on clinician opinion of benefit.

2.4.1.2.5 Clinician perspective: patient scenario comparison.

This method has been applied to a visual analog scale (VAS) of pain measurement by providing clinicians with a series of descriptions of patient scenarios.[67, 72] The clinicians provided their assessment of each hypothetical patient's pain on the VAS and then also rated how much different the current patient's pain was compared to the previous scenario, using descriptors such as "much less/more", "a little less/more", etc.. The MID was taken as the difference in the VAS between pairs associated with a clinician difference rating of "a little less/more".

2.4.1.2.6 Clinician perspective: prognostic rating scale.

A prognostic rating scale method was used as the criterion for within-person change for the Neck Disability Index (NDI).[73] A clinician-based assessment of prognosis from soft-tissue neck injury (ranging from little or no change expected to excellent improvement expected) was used to determine an important change in patients. A receiver operator characteristic curve was then applied to determine the NDI change score that best discriminated between those who had important improvement versus those who did not.

2.4.1.2.7 Data driven approach.

Wyrwich et al. have reported on an approach to defining the MID that is based strictly on the distribution of the instrument scores.[74] They suggest an entity which they call the standard error of measurement (StErMe) is, essentially, a proxy for the MID and is calculated as:

$$\text{StErMe} = SD * \sqrt{(1 - \alpha)}$$

where SD = the standard deviation of the scores at baseline and α = Cronbach's alpha coefficient. They found that their approach resulted in MID estimates that were similar to the patient global rating approach.

Another data driven approach uses the *effect size* to determine the MID.[75] The effect size provides a standardized measure of differences in a health status instrument. Although there are different ways in which it has been defined mathematically, one definition is the difference between the mean scores of two groups divided by the standard deviation of the scores at baseline.[76] The effect size can then be called "small" (or the MID) if it is 0.20, "moderate" if it is 0.50, or "large" if it is greater than 0.80.[75, 76]

2.4.1.2.8 Discerning important improvement: improvement criteria.

This approach involved surveying clinicians and providing them with data about several randomly selected patients from clinical trials using a certain outcome measure[77] The clinicians were provided with baseline and end-of-study scores and asked to assess whether they felt the patient had improved. An important change in the outcome measure was that which the vast majority of clinicians agreed had improved.

2.4.1.2.9 Discerning important improvement: achieving treatment goals.

In determining improvement of low-back pain with physiotherapy, Riddle et al. used the achievement of pre-determined treatment goals as the criteria for a clinically important improvement.[78] Prior to initiating treatment, specific treatment goals were set by the patient and clinician. Various other change scores were then compared to this criterion measure. This method looked at within-patient changes and noted that the threshold for important improvement varied with the baseline status of the patient.

2.4.2 Problems with the MID

A careful examination of the previously quoted definition by Jaeshcke et al. will to serve to highlight the currents points of debate regarding the MID and its potential limitations.[66] It is recognized at the outset of this discussion, that this definition does not represent a consensus opinion. Nonetheless, it provides a fruitful basis for a critique of the MID concept.

The first point of controversy is the frame of reference for determining an important difference. In Jaeshcke et al.'s definition, which was referring specifically to measures of quality of life, it is the patient's perception of the difference as "beneficial" that matters.[66] While it may be a general point of agreement that the patient's perception of benefit and meaningful change is paramount, one might also argue that, especially for measures of health status rather than quality of life, physicians should also be used as a frame of reference. In fact, as discussed in the previous section, physicians have been used by some researchers to determine the MID.[67]

Another point of controversy, and perhaps the most fundamentally important to the concept of MID, is that the MID definition includes mention of the "absence of troublesome side effects and excessive costs" of a certain management option.[66] This is an explicit acknowledgement that the MID is not an inherent property of a health status instrument, but, rather, is very much dependent on the particulars of a situation. For example, in the case of a health status instrument for angina, the MID would be different when the "change in the patient's management" was a medication with a low side-effect profile rather than a major, life-threatening operation. While this is obvious, it casts serious doubt on the tenability of the MID as an absolute entity.[79] If, in fact, the MID is not an absolute entity, then of what use is it at all? Even ignoring this fundamental question, one notices that, of the currently described methods for determining the MID (see previous section), few explicitly deal with the issue of the relative risks of the management options for the patient.[67] In other words, most assumed the MID to be an absolute entity. Related to this is the implicit assumption in the MID definition that the response options to a health status instrument are, in fact, interval in nature. This assumption underlies much of the work in this field, so the MID can not be singled out for relying on this unproven assumption. However, the truth or lack of truth of this has, perhaps, the greatest conceptual impact on the MID.[80, 81] This is because, in its attempt to attribute interpretability to a numerical scoring system, it is assumed that the difference in a score of "50" compared to "75" is exactly the same as the difference between "25" and "50". Of even greater difficulty is the assumption that a positive change in the numerical score carries the exact same magnitude of clinical importance as an equal negative change. Intuitively, one would likely question this on face value: it would seem to make sense that an improvement in health would be perceived quite differently by patients than a deterioration in health, as has been suggested empirically.[82] If this is the case, then, once again, the MID loses its claim as an absolute entity.

Finally, an inherent limitation of the MID is that it only addresses the

interpretability of a difference in scores. It ignores the equally important question of how to interpret an absolute numerical score. This is important in purely descriptive cross-sectional studies of a given population.

2.4.3 Alternative to the MID

If the MID is not an absolute entity, but, rather, is dependent on the context of its measurement, the initial state of the patient's health, whether the difference in health involves an improvement or a deterioration, then it loses much of its value to researchers and clinicians. In the absence of the MID, are there other methods to address the important, and admittedly difficult, task of ascribing interpretability to the numerical score of a health status measure? In an earlier work, this author suggested a method which has since been implemented, quite coincidentally, by another group, and will be discussed here.[1, 83]

The basic elements of this method involve translating a numerical score on a health status measure to its equivalent score on a separate health status measure for which health utility values have already been described. This effectively translates the numerical scores into a health utility score. Utility scores, which define a "preference for or a desirability of a particular outcome",[84] have a more tangible, economical meaning and have generally gained acceptance, or, at least, familiarity, amongst clinical researchers. Utility scores are perhaps more inherently interpretable to clinicians because they are based on absolute reference points, usually "perfect health" and "death".[85] It should be noted, however, that there is some evidence to suggest that these reference points may not be so absolute to all individuals.[86] Nonetheless, utility scores also have the advantage of allowing for some meaningful comparisons across groups and across different health status measures, especially because utility scores are usually created on an interval scale.[87, 88] Such comparisons can then be used to perform economic evaluations.[89] However, in order for this method to be successful, certain criteria need to be met, similar to those outlined earlier by Guyatt.[65]

First, in order to speak meaningfully of health utility scores, the health status measure of interest should be a general and comprehensive measure of health, rather than of a very specific aspect of health. Otherwise, the range of scores could ignore large and relevant components of health and could not fairly be said to represent the patient's overall health status.

Second, a useful, established general measure of health, with predetermined utility scores, must be found and must also be applicable to the population of interest. An excellent example of this is the Health Utilities Index-2 (HUI-2).[90] This is a multi-attribute health classification system that has demonstrated good psychometric properties in various pediatric populations.[91-93] The HUI-2 assesses a child's health status in the following attributes: sensation, mobility, emotion, cognition, self-care, and pain. Within each of these attributes, the child is assigned to a particular level of functioning and from this an overall health utility score can be calculated.

Third, scores from the health status measure must be translated into HUI-2 scores. This requires that there be a relationship between the two scores, such that some form of mathematical transformation will consistently convert one to the other. This relationship could be of any type, but it would certainly be preferable if a simple linear relationship existed between the two measures. In the case of a simple linear relationship, it is important that a strong correlation exist. Guyatt has suggested that in order to use an external standard for comparison, a correlation of at least 0.5 is needed.[65] The need for a strong correlation between the psychometric measure and the utility measure might be the most difficult criteria to meet. Revicki and Kaplan, upon reviewing the literature, found that the correlation between such measures was usually only low or moderate, resulting in regression equations with less than 50% explained variance.[94] Furthermore, even if a strong mathematical relationship could be demonstrated, one must always be aware of the conceptual leap required in transforming from a simple measure of health status to a measure of health status *preference*.

As mentioned previously, a very similar approach was recently taken by Nichol et al. in order to derive health utility scores for the SF-36.[83] Using a large sample of patients (n=6921) studied as part of a managed-care pharmaceutical care practice research project, patients completed both the SF-36 and HUI-2. A mathematical linear model was constructed using the HUI-2 utility score as the dependent variable and each of the 8 domains of the SF-36, along with patient age, as the independent variables. The correlation between the domain scores of the SF-36 and domain scores of the HUI-2 ranged from less than 0.20 to as high as 0.70. Each of the individual SF-36 domains demonstrated a correlation roughly in the 0.50 range with the HUI-2 utility score. The final linear model had a R^2 of 0.51.

The advantages of this method have been described and, if the previous criteria can be met, then this will be the preferred methodology for establishing the interpretability of the hydrocephalus outcome questionnaire.

2.5 Structural Equation Modeling

Structural equation modeling (SEM) is a form of confirmatory factor analysis; a "comprehensive statistical approach to testing hypotheses about relations among observed and latent variables."[95] These hypothesized relations are usually represented in a diagrammatic form. The hypothesized model is then tested against the available data to assess how well it fits. A brief review of the technique is presented here.

The terminology of SEM is at times unclear. For the purposes of this work the following terminology will be followed, as per McDonald and Ho's recent review of the literature.[96] One of the first steps in SEM involves defining the model. The *measurement model* represents the relationship of the observed variables as indicators of the unobserved, latent variables (or *factors*). Usually, an observed, indicator variable loads onto only one factor. The *path model* describes "relations of dependency – usually

accepted to be in some sense causal – between the latent variables."[96] These relationships may be represented by *directed arcs*, which imply a one-way cause-effect relation between two factors, or *nondirected arcs*, which imply some correlation between factors due to common causes that are not accounted for in the model. Graphically, directed arcs are represented by single-headed arrows and nondirected arcs by double-headed arrows. The combined path model and measurement model makes up the *structural model*.

In SEM parameters can be either *fixed*, *constrained*, or *free*.[97] A *fixed* parameter is one that is fixed to a pre-specified value (usually 0 or 1). A *constrained* parameter is one that is set to be equal to another parameter, but the exact value of this is not pre-specified. A *free* parameter is one whose value is unknown and not constrained. There can be several *free parameters* and one of the goals of the exercise may be to provide estimates of these parameters. One method of deriving these parameters is to use an iterative method. This method begins with start values for the free parameters and constructs a covariance matrix based on these values. This is called the *implied covariance matrix*. This is compared to the *observed covariance matrix*, derived from the data set. The difference between these results in the creation of a *residual matrix*. This is performed for each iteration (with each iteration assuming different parameter values) and the residual matrix is compared to the previous iteration until it is minimized, i.e., until the implied and observed covariance matrices are as close as possible. At this point, the iterative procedure is said to have converged.

At this point, some form of statistical evaluation of the fit of the specified model is carried out. In general, this can be one of two forms of evaluations: a measure of absolute fit or a measure of relative fit. The most common measure of absolute fit is the χ^2 goodness-of-fit measure. This involves calculating a χ^2 statistic based on the final residual matrix and the sample size.[95] The smaller the statistic, the closer the fit, with zero implying a perfect fit between the implied and observed covariance matrices. The major limitation of the χ^2 goodness-of-fit test is that, for it to be valid, two assumptions must be met: the data must have a multivariate normal distribution and the specified model must be the correct one.[95] It would be rare in clinical research that both these assumptions would be met, casting doubt on the validity of the χ^2 statistic. Largely for this reason, indices of relative fit have become popular. Indices of relative fit do not compare the implied and observed covariance matrices. Rather, they compare the fit of the specified model to the fit of a null model. A null model is one in which no specific relations among the observed variables are specified, i.e., there are no factors. Indices of relative fit usually vary within a metric that ranges from 0 to 1.0, with a higher value representing a better relative fit. The major limitation with these indices is that their interpretation is quite subjective. The cut-off value above which the specified model might be considered a "good fit" is debatable, although many would suggest a level of at least 0.90.[95, 98]

2.6 Qualitative Research

Methods borrowed from qualitative research were used in the Item Generation phase of this project and, therefore, a brief discussion of this topic is presented. Qualitative research is usually defined by noting its differences from quantitative research.[99] In most general terms, it has been said that qualitative research deals with words rather than numbers. A more fundamental difference, however, is the inductive nature of the research, with theory emerging from the research. But qualitative research itself is not homogeneous: there are different research traditions in qualitative research and each is rooted in a slightly different philosophical perspective of the research world.[100] That is, the ontology (the nature of reality), the epistemology (the knowledge of reality), and the methodology (the ways of studying reality) differ to varying degrees. Some of the more important research traditions are: phenomenology, hermeneutics, grounded theory, and ethnography.

The methods of doing qualitative research are also varied. Examples include:[99]

i. Participant observation
ii. Qualitative interviewing.
iii. Focus groups.
iv. Analysis of discourse and conversations.
v. Analysis of texts and documents.

For this project, the focus group method was used to gather options from parents of children with hydrocephalus about their children's health and well-being.

2.6.1 Focus Groups

The focus group method involves a group interview in which "there are several participants (in addition to the moderator/facilitator); there is an emphasis in questioning on a particular fairly tightly defined topic; and the accent is upon interaction within the group and the joint construction of meaning."[99] This last element highlights one of the major advantages in the use of focus groups: the answers people give in this setting may be different than in one-to-one interviews. A person may hear someone else's response and then be encouraged to say something that he/she otherwise would not have thought of or said. Similarly, participants may challenge each other's responses in a way that might not seem appropriate for an interviewer to do. As well, because the role of the moderator should be quite minimal, the participants of the focus group are able to bring to the fore the issues that they feel are most relevant.

There are, however, some disadvantages to the focus group method, as well. A major problem is the possibility of deleterious group effects.[99] This includes, for example, one or two participants who dominate the discussion, at the expense of other participants' opinions. In this situation the role of the moderator becomes important, to encourage the quieter participants to voice their options. Another potential disadvantage of focus groups is that it may inhibit discussion on topics that participants feel are too personal or embarrassing.[101] Therefore, one must be sure that the participants feel

comfortable discussing the research topic of interest in a group setting. Otherwise, individual interviews may be preferable. There are also some practical issues to overcome in conducting focus groups, including finding and scheduling a time/place that is convenient for several people and the problem of recording and transcribing the session (see next section).

2.6.2 Conducting Focus Group Research

There are numerous technical points that need to be addressed when conducting focus group research. Many of these points require decisions of judgment on behalf of the investigator, since hard-and-fast rules rarely apply.

The optimal size of a focus group is controversial. Although typically suggested numbers range from 6 to 10, many reports of focus groups in the literature actually use fewer people than this (more often in the 4 to 5 person range).[99] One of the reasons for this is the problem of people who agree to participate but who then do not show up: the "no-shows". While this is to some degree beyond the control of the investigator, it perhaps can be minimized by making personalized telephone contact with participants prior to the focus group and gauging their true level of interest. Another factor to consider in focus group size, as suggested by Morgan, is the research topic itself.[102] That is, if the topic is one for which it is anticipated that the participants will have a lot to say, then smaller groups are recommended. This is particularly so if it is a complex topic or one which emotionally involves the participants. Such would be the case, for example, in this project's topic: the health status and well-being of the participants' children. Another advantage of smaller group size is that some participants who might be hesitant to speak out in a group setting, will likely be less intimidated in a smaller group. This might encourage them to participate to a greater degree than might have been the case in a larger group setting.

The optimal number of focus groups that need to be conducted also depends on what is being studied. The most common criteria used to determine when enough groups have been sampled is that of *saturation*.[99, 103, 104] That is, one "continue(s) until comments and patterns begin to repeat and little new material (is) generated."[105] However, the number of groups must also allow for adequate sampling of the population of interest, such that representation from all potentially relevant individual profiles is achieved. For this project, this would mean that representation of children with varying underlying etiologies and varying ages is required.

The role of the moderator should be a non-intrusive one, serving mostly as a facilitator of discussion.[99, 101] The moderator should raise open-ended questions relevant to the topic of study and re-direct any discussion that has gone wildly off-topic.[106] As well, it is very important during the focus groups that there be some trustworthy method of transcribing what is said.[107] While this may be as simple as the moderator transcribing in writing, it is usually better to have some form of recording device to allow for an accurate and comprehensive record of everything that is said and by whom. Probably the best way of achieving this is by video-taping the session. This will not only record the

verbal content, but it also makes it very clear who is speaking. Whatever type of record is made, it then needs to be transcribed such that its information is put into written form for whatever type of analysis is needed for the project.

2.7 Mail Surveys

The heart of this project is the development of a questionnaire which will be administered both in person and by mail. Considerable empirical research has been compiled about the features of questionnaire administration, particularly mailed questionnaires, that result in improved response rates and responder interest. Some of these features will be reviewed here, because they were used throughout various stages of this research.

2.7.1 Techniques to Increase Mail Survey Response Rates

2.7.1.1 Prenotification

Prenotification involves alerting potential respondents that they will be receiving a survey shortly.[108] This may be done by telephone or mail, although there might be an advantage to using telephone contact, since it adds an element of personalization. The impact on response rate of prenotification has been noted to be about equivalent to that of a single post-mailing reminder.[108, 109]

2.7.1.2 Cover Letter

The cover letter is a very important component of any mailed survey.[110] It serves to introduce the research and the investigators. This letter should be aimed at the target sample and explain the rationale for the study in a way which will encourage participation and interest. As well, if a study is being sponsored by a university or hospital, it is important to display this prominently to establish credibility and trust amongst the respondents.[108] Contact information should also be provided so that if the respondent has any questions or concerns, they know where to turn. As well, the letter should provide some indication of what time-frame is expected for completion of the survey. In general, it is recommended that a suggested duration of time be given rather than a strict deadline date. This is believed to increase response, since in the presence of a strict deadline date, those who have not completed the survey will give up entirely thinking that it is too late for them to respond.

A more personalized letter may also help response rates, although the effect may be small.[108] Personalization can take different forms. For example, the salutation of the cover letter might mention the person's name rather than something more general or anonymous. As well, the letter may be personally signed by the investigator, rather than using a stamp or no signature at all.

2.7.1.3 Questionnaire Format

Formatting of the questionnaire is important with the primary goals being easy comprehensibility and easy completion. The best way to accomplish this is by pre-testing the questionnaire with a sample representative of the population of interest. This will allow for early warnings about lack of clarity and other problems that might be encountered that could reduce response rate.

2.7.1.4 Mailing Techniques

There may be some effect on response rate of the type and appearance of postage used to deliver a mail survey.[108] For example, a slight advantage in the use of stamps rather than metered mail has been shown by some.[111] Presumably, the respondents are less apt to consider a stamped envelope "junk mail". A similar advantage has been shown for stamps placed on the return envelopes.[111] Respondents "do not want to waste the stamp by not returning the questionnaire and yet are not crass enough to peel it off and use it for their own purposes."[108]

2.7.1.5 Incentives

Providing the respondents some form of incentive to complete the questionnaire usually results in improved response rate.[108] The most direct form of incentive is to give money: either prepaying the respondents or promising money after completion of the questionnaire. Usually, the latter method is less successful than the former. What is perhaps most interesting about the effect of monetary reward on response rate is how the amount of reward correlates with improved responses. Most studies suggest that this is not a simple relationship. Rather, while there is some improved response with increasing rewards, this may level off at a certain level of compensation. Mangione provides some interesting postulates as to why this might be the case.[108] For example, at a low level of reward, say under $2, it is clear that the reward is not meant as fair market compensation for the completion of lengthy questionnaire, but, rather, a token of good faith. With greater compensation, however, the reward begins to approach fair market value for their time and respondents may view this as a "job" which they are free to decline. Another theory is that a large reward may give the impression that there is something wrong with the study, while a smaller reward might suggest to them the importance of the study. Of course, these are all just postulates that have not been tested.

2.8 Summary

This chapter dealt with some rather diverse methodological issues. The goal was to lay some groundwork for the practical work to follow. As noted at the beginning of this chapter, some methodological issues were not presented here. These have either been discussed by the author in previous work[1] or will be dealt with in their appropriate sections later in this work.

The remainder of this thesis will deal with the practical aspects of developing a new health status outcome measure for children with hydrocephalus. Separate chapters will be devoted to each of the various stages of the project: Concept Development, Item Generation, Item Reduction, Reliability Testing, Validity Testing, and Interpretability.

3. The Concept and the Population

3.1 Concept Development

The first step in developing a new measurement instrument is to define what it is you wish to measure. For some things in medicine, the task is quite clear. For example, there are standard definitions that one could rely upon if one were interested in measuring blood pressure or some specific laboratory test value (although the details of measurement may vary). However, when the object of measurement is less well defined, the task is more difficult. In general terms, the goal of this instrument is to measure the health status of children with hydrocephalus. While this statement is simple enough, the definition of health, or rather, how one chooses to define it, will impact on both the type of instrument one might develop and what that instrument actually measures. This chapter will review some of the attempts in defining health itself, followed by an approach to developing the concept of health status for the purposes of this new instrument.

3.1.1 Definitions of Health

There are several definitions of health that have been put forward and no one definition seems to stand above the others as absolute. Rather, each particular definition appears best suited for use in different situations.

3.1.1.1 World Health Organization Definitions

The World Health Organization (WHO) is probably the most referred to source for definitions for health worldwide. One of the most quoted definitions is their revised 1984 definition of health:[112]

> The extent to which an individual or group is able to realize aspirations and satisfy needs, and to change or cope with the environment. Health is a resource for everyday life, not the objective of living; it is a positive concept, emphasizing social and personal resources as well as physical capabilities.

Of course this definition is very far reaching and has implications beyond what would likely be of interest to most clinical researchers. This WHO concept of health relates not only to medical health, as is traditionally viewed by clinicians, but also to issues of social health of the individual and the environment in which they live. This implies a large degree of political influence on the definition of health. That is, in an oppressive regime,

for example, the ability of an individual to "realize aspirations and satisfy needs" may be compromised, despite adequate medical health. While the WHO definition of health is important for the geopolitical determination of global health, it is too broad for clinical research use.

An earlier WHO definition of health defined it as "a state of complete physical, mental and social well being and not merely the absence of disease or infirmity."[112] This is probably the best known definition. Saracci placed this definition in its historical context:[113]

> This by now classical definition of health, conceived in the aftermath of the second world war, when peace and health were seen as inseparable, had one merit lasting long beyond the circumstances of origin: it made explicit that disease and infirmity, when isolated from subjective experience, are inadequate to qualify health.

That the determination of health could not ignore the social and psychological aspects was, at the time, a revolutionary idea. Over the years, this has become a generally well-accepted tenet in defining health. However, Saracci argues that this WHO definition is not a definition of health, but, rather, a definition of happiness.[113] The distinction, he argues, is relevant, since *health* can be seen as a positive and universal human right, but not *happiness*: it "cannot be delivered or imposed on a person by any societal action. Happiness is strictly subjective both as an achievement and an appreciation."[113] Including happiness in the definition of health would have substantial resource-allocation implications, since the quest for happiness "is essentially boundless".[113] Similar criticisms have been put forth by others.[114]

Despite criticisms of their rather broad nature, the WHO has made attempts to operationalize these definitions of health. Their first attempt was the 1980 International Classification of Impairments, Disabilities, and Handicaps (ICIDH).[115] One of the main goals of this was to be able to measure the disability prevalence in populations across the globe. The approach taken was one of trying to classify the *consequences of disease*.[116] During the development phase of this classification, it was decided that at least two axes needed to be considered: that of the individual and that of the interaction between the individual and the environment. Expanding on this initial premise led to a three level classification:

i. Impairment – corresponding roughly to organ-level dysfunction
ii. Disability – corresponding to dysfunction in individual action
iii. Handicap – corresponding to dysfunction in societal interaction

The ICIDH has been translated into many languages and has been very widely used around the world for numerous types of assessments ranging from community-level prevalence studies to individual patient evaluations of treatment.[116] However, the definitions of the various terms (i.e, impairment/disability/handicap), the possibility of overlap in classification, and the ignored cultural context of handicap have led to some

criticism of the ICIDH.[116, 117]

Another problem with the ICIDH was discussed by Susser in an interesting exposition on the ethical and philosophical aspects of defining health.[114] He argued that health is related to the concept of normality and that this, in turn, involves at least three potential interpretations. Normality can be considered in a dichotomous pathological sense: the disease is either present or absent. Normality can also be considered in a statistical sense, based on the distribution of a given attribute (or levels of an attribute) across a population. Finally, normality can be considered in a social sense, defined by societal values. Social values and social structure can play a large part in the "lack of correspondence between the existence of organic disease...and of the sick role."[114] This became one of the main criticisms of the ICIDH – that it implied a causal link from impairment to disability to handicap, thereby ignoring other potentially relevant factors.

More recently, in an attempt to deal with some of the criticisms of the ICIDH, the WHO produced a revised health classification scheme called the International Classification of Functioning, Disability and Health (ICF).[118] The goal of the ICF is to provide a "meaningful picture of the health of people or populations, which can then be used for decision-making purposes."[118] The ICF consists of the following parts:

Part 1. Functioning and Disability
- (a) Body Functions and Structures
- (b) Activities and Participation

Part 2. Contextual Factors
- (c) Environmental Factors
- (d) Personal Factors

Within each component there are domains and categories with positive and negative attributes. For example, some of the domains within the Activities and Participation component include: communication, mobility, self-care, interpersonal interactions and community/social/civil life.[118] The ICF is regarded by the WHO as a *components of health* classification, rather than a *consequences of disease* classification (such as the ICIDH):[118]

> "Components of health" identifies the constituents of health, whereas "consequences" focuses on the impacts of diseases or other health conditions that may follow as a result. Thus, ICF takes a neutral stand with regard to etiology so that researchers can draw causal inferences using appropriate scientific methods."

In this way, this new classification of health avoided the implication of a causal link connecting impairment, disability, and handicap.

3.1.1.2 Other Definitions of Health

It is clear from the WHO definitions that health is almost certainly a

multidimensional concept. There can be, however, some disagreement as to what these dimensions (or domains) are. In 1981, Keller reviewed the literature in the area of defining health and ranked the most commonly cited domains.[119] Not surprisingly, the most common was physical/biological health, followed by emotional health and social health. Less commonly cited dimensions included spiritual health, relationships, and cultural health. As was the problem with the WHO definition, the inclusion of such dimensions increases the scope of health beyond that which would be considered useful to most medical researchers.

Patrick et al. attempted to provide a more restricted, operational definition of health, with emphasis on the practicality of measurement.[120] They viewed health as a composite of an individual's level of function at a point in time together with their prognosis for change in function:[120]

> Precisely stated, health status is the product of the social preferences assigned to levels of function and the probabilities of transition among the levels over the life expectancy of an individual.

The distinction between function and prognosis, they argued, was important because these dimensions were traditionally confused.[120] They give the example of two individuals confined to bed due to gastrointestinal disease: one due to transient diarrhea and the other due to an incurable malignancy. While their level of function might be similar at a given point in time, their prognosis certainly differs, and, therefore, Patrick et al. argue, so does their health status.

Within the context of children's health status, numerous authors have provided varying examples of conceptual frameworks. These works include those of Bergner, Eisen et al., Feeny et al. and Cadman et al., and Parkin et al..[59, 91, 121-123] These different works tend to have recurring themes in their conceptual framework for health. Borrowing from these themes, a preliminary multidimensional structure of health for the new hydrocephalus instrument follows:

Overall Health

- → **Physical Health**
- → **Social Health**
- → **Emotional Health**
- → **Cognitive Health**
- → **Level of Independence/Self-Care**
- → **Pain**
- → **Communication**
- → **School Performance**
- → **Future Development/Prognosis**

This framework formed the basis for proceeding with the Item Generation phase of instrument development. It was hoped that during that phase of the research the concept would become better defined, based on input from multiple sources.

3.2 Defining the Population and the Respondents[§]

The inclusion criteria for this project (and, therefore, the target population for the new outcome measure) and the rationale for using parents as the primary respondents for the questionnaire have been discussed in detail in a previous work.[1] This will be repeated here for ease of reference.

3.2.1 The Population

The following inclusion criteria will be used to define the population of interest:

3.2.1.1 Diagnosis of hydrocephalus

This requires a diagnosis of symptomatic hydrocephalus, at some point in the child's history, based on clinical and radiographic criteria, that have been well-established in the literature. The clinical symptoms and signs include:

Symptoms: headache, nausea, vomiting, decreased level of consciousness, irritability, decreased school performance, loss of developmental milestones.

Signs: papilledema, bulging fontanelle, nuchal rigidity, 6^{th} nerve paresis, loss of upward gaze, new or increased seizures, increased head circumference.

Radiographically, there should be evidence that, at some point, the child had enlarged cerebral ventricles on CT or MR. Since this instrument is aimed at children with chronic hydrocephalus, the diagnosis must have been present for at least 6 months.

This will include all children with the diagnosis, whether or not they have received surgical treatment for their hydrocephalus. Hydrocephalus is the sequela of a great many different primary conditions. None will be specifically excluded.

3.2.1.2 Ages 5 to 18 years old

Hydrocephalus affects children of all ages and an instrument that captures a wide a range of ages is preferable. However, because health problems can change with age, the extent of age ranges covered must be tempered by the desire for a single, simple instrument that is relevant to the age ranges being considered. For this study, the instrument will deal only with children in the 5 to 18 year old age group. The lower limit of age was decided because the health issues for younger children, i.e., preschool age, will

---

[§] This section is taken largely from the candidate's Masters thesis[1] and is not being presented here as original new work. It is provided simply as background reading for the remainder of this Doctoral thesis.

be very different and questions about school will not be relevant to them. However, in order not to be premature about excluding this group, younger children will initially be included in the early stages of this research project. If the items on the final questionnaire appear, to the parents of these younger children, to be not relevant, then this group will be excluded from later parts of the study. The upper age limit represents a fairly well-accepted limit to the pediatric age group and, as a matter of practicality, also represents the upper limit of age of patients allowed to be treated at Hospital for Sick Children.

Despite limiting the ages to this range, it is quite possible that health issues within this range will differ. Other pediatric health instruments have subdivided their population into smaller groups, with separate questionnaires for each group. For example, Parkin et al. divided their population of spina bifida patients into those under 12 years and those over 12 years.[59] Those over 12 completed their own questionnaires, while the parents were the respondents for the younger group. As well, in a review of quality of life instruments in children, similar age cutoffs were found in many other pediatric instruments.[124] The advantage to using separate questionnaires is that it allows for a very specific assessment of health issues relevant to that age group. However, the disadvantage is that it does not allow for a simple comparison of scores between the age groups. This further complicates attempts at comparing outcomes between groups of patients which might be composed of children of varying ages. This would defeat one of the main purposes of developing this new health outcome measure. Therefore, for this new instrument, a single questionnaire will be developed to cover the entire age range from 5 to 18 years.

3.2.2 The Respondents

The children who are the focus of this instrument represent a range of ages. There are certain limitations in approaching the children as the respondents. The primary issue is their ability to answer questions about their health status in a meaningful way. Pantell and Lewis point to certain response biases that tend to be greater in children, such as, position biases (e.g., tendency to choose the first answer), acquiescence response bias, limited understanding of negatively worded items, and time perception differences.[125] Previous work has shown that many children as young as 7 years old do appear to possess the minimum abilities to answer such questions in a reliable fashion.[126] However, this was determined on a group of children with asthma, who were otherwise cognitively normal for their age. The population of children with hydrocephalus has varying degrees of cognitive impairment and it can not be assumed that all hydrocephalic children above 7 years old would be capable of answering questions about health status. A more similar population was that of Parkin et al., in which children with myelomeningocele (many with hydrocephalus as well) older than 5 years of age were the subject of a quality of life instrument.[59] They divided their population into two age groups: those 5-12 years old and those 13- 18 years old. It was found that many of the children in the younger age group required the help of their parents to complete their questionnaire. While parents were instructed to answer questions from the child's point of view, it is probably difficult to

separate the influence of parental perception from the child's perception. This brings into question whether questionnaires filled out exclusively by children are measuring the concept in the same way as those filled out by parental proxy. If not, how will this then effect the measurement?

Because it is anticipated that a significant proportion of children will require parental assistance or even full parental proxy response, the effect may well be that the instrument ends up measuring different forms of the concept, depending on who is responding. The interpretation of such data then becomes very complicated. It is reasonable to conclude that by measuring different forms of the concept, the reliability of the instrument suffers as well. This would tend to obscure any true effects that may exist when comparing different interventions. This is clearly not an optimal situation.

To account for this, parents will be used as the respondents in all cases for this new instrument. In effect, this is the lowest common denominator of respondents, since in all but the rarest cases, parents will be able to respond on behalf of their children. This creates a more homogenous environment and will allow, hopefully, for measurement of the concept in a more uniform way.

3.3 Summary

This chapter presented the theoretical and practical justifications for the chosen concept, population, and respondents. The next stage of the project is Item Generation, in which an attempt will be made to collect a comprehensive list of items for inclusion in the final questionnaire.

4. Item Generation

The major focus of item generation was to create a comprehensive list of items that represents the health status of children with hydrocephalus. This involved the use of multiple informants, most prominently the parents of the children themselves. The techniques used to extract this information from the informants were borrowed from the field of qualitative research.

4.1 Methods of Item Generation

Three sources for item generation were used:

4.1.1 Expert Sources

Doctors and nurses with experience in dealing with children with hydrocephalus were approached to provide items for the instrument. A brief, open-ended questionnaire with an instruction page was provided to four pediatric neurosurgeons and two clinical pediatric neurosurgical nurses from the Hospital for Sick Children, Toronto. (see **Appendix A – Information Sheet for Item Generation from Expert Sources**) These were given to each participant directly by the investigator with instructions to complete as soon as possible. After each had completed the questionnaire, the investigator performed an informal, semi-structured interview with the participant. The purpose of this was to clarify the items that they had written and to elicit any further items they had thought of since having completed the questionnaire. These interviews were conducted approximately one week after completion of the questionnaire.

4.1.2 Literature Review

The literature was reviewed to identify previous work in the health status of children in general and those with hydrocephalus, specifically. Items used in these studies were screened by the investigator, searching for those that were felt to be useful for the new instrument. From this, the following were used as sources of items:

Scale of Perceived Self-Competence, Harter, 1982[127]
HRQL for Rhinoconjunctivitis, Juniper, 1998[128]
Self-esteem in Myelomeningocele, Wolman, 1994[129]
Functional Status Revised-II, Stein, 1990[130]
CHIP, Starfield, 1993[131]
Multiattribute Health Classification System, Feeny, 1992[91]

Warwick CHMP, Spencer, 1996[132]
Headache Quality of Life, Langeveld, 1996[133]
CHAS, Budd, 1994[134]
Impact-on-Family Scale in Myelomeningocele, McCormick, 1986[135]
Quality of Life in Myelomeningocele, Parkin, 1997[59]
Quality of Life in Epilepsy, Hoare, 1995[136]
PCQL-32, Varni, 1998[58]
Outcome in Myelomeningocele, Kalucy, 1996[137]
Rutter Scale, Goodman, 1994[138]

4.1.3 Parent Focus Groups

In order to generate items from the parents of children with hydrocephalus, it was decided to speak to them in a focus group setting. Discussion about some of the theoretical and practical aspects of focus groups have been discussed elsewhere in this work (see **Section 2.6**).

4.1.3.1 Selection of Sample

The Hydrocephalus Database at the Hospital for Sick Children was used to randomly select patients with the following characteristics:

- between 2 to 18 years of age
- diagnosed with hydrocephalus for at least 6 months
- current address within the greater Toronto area

Initially, a list of over 100 patients was generated which was subdivided on the basis of underlying etiology and age. From these subdivided lists, 32 parents were selected in random order, in order to provide good representation of all age ranges and etiologies. This was done in an iterative fashion, such that after each focus group another set of parents were called (32 in total), until saturation in new item generation was reached.

Attempts were made to call these 32 parents at various times during the day and evening. Five could not be reached (1 phone number was not in service, 2 did not speak English, 1 was out of town for an extended period, and 1 no answer despite repeated attempts at calling). Amongst the 27 parents who were reached, three were not interested in participating in the focus groups. Ten were interested but did not participate because of the timing of the focus groups. Fourteen sets of parents did participate in the focus group discussions. These parents represented fourteen children with the following characteristics:

TABLE 4.1. Characteristics of focus group participants

Age

Mean:	**7.5 years**	**(min: 2.3 to max: 16.8 years)**	
Distribution:	**< 5 years**	**=**	**4**
	5-7 years	**=**	**6**
	8-12 years	**=**	**2**
	13-18 years	**=**	**2**

Etiologies

- **aqueductal stenosis/ congenital** = **6**
- **myelomeningocele** = **2**
- **other** = **6**
  - **intraventricular hemorrhage (1)**
  - **meningitis (3)**
  - **head injury (1)**
  - **tumour (1)**

4.1.3.2 Focus group procedure

A total of five separate focus groups were held on weekend afternoons in conference rooms at the Hospital for Sick Children. Each focus group consisted of between two to four sets of parents (between two to five persons). Whenever possible, an attempt was made to provide a diverse mix of etiologies at each focus group. Coffee and refreshments were served during the sessions.

The sessions were facilitated by the investigator (who had not been involved in the medical care of any of the children and had not previously met the parents). Each session began with a brief description of the purpose of the research project in general and the session in particular. The parents were informed of the presence of a video camera which recorded the entire discussion. An outline of the various topic headings that were to be covered during the session was provided to the parents. The parents were then asked to read and sign a consent form.(see **Appendix B – Sample Consent Form for Focus Group Participation**)

The facilitator began each session by asking the parents if there was any particular topic heading they would like to begin with. The parents were encouraged to bring up any health issues that were relevant to their children. The facilitator maintained a very discrete role, intervening only to help cover topic areas that were missed or to encourage parents who were relatively silent to have their say. The facilitator simultaneously wrote down the issues on a chalkboard as the parents brought them up.

Once all the parents were satisfied that all issues they had had been discussed and all topic areas were covered, the parents were shown a list of the items generated from the expert sources. After reading over the list, the parents were asked if they could think of

any further issues to discuss. This occasionally led to further discussion.

Once all discussion was completed the parents were thanked for their time and given free parking pass vouchers. The sessions lasted a total of between 90 minutes to 2 hours.

After each session, the videotapes were fully reviewed and all issues that were mentioned were transcribed by the investigator.

4.1.4 Child Informant

In order to augment the information provided by the parents, each parent who participated in the focus group was asked if they would allow their child, if agreeable and at least 12 years of age, to participate in a semi-structured interview. Although 2 parents agreed to present this option to their children, only 1 child agreed (a 16 year old female with spina bifida). This child was interviewed by the investigator in a semi-structured format which followed the general topics of discussion in the parent focus groups. Although the session was also videotaped, the camera view did not include the child in the screen shot at her request. However, the entire audio of the interview was captured by the videotape. The interview lasted approximately 40 minutes and despite in-depth discussion of the topic areas, no new items were generated from this process. No further interviews with children were pursued after this.

4.2 Results of Item Generation

An initial comprehensive list of 274 items was generated from all three sources. The following is the breakdown of number of items from each source:

Expert Sources..... 43 items
Literature Review..... 54 items
Focus Groups..... 177 items

After eliminating clearly redundant items, 187 items remained. These are listed in **Appendix C. Preliminary List of Items Generated During First Phase of Research.** These appeared to be broad in their coverage of all major areas of health.

4.3 Discussion

The item generation phase of this project was important in that it needed to be comprehensive. Upon entering the next stages of research, there would be no turning back and the opportunity to include missed items would be lost. Therefore, the approach that was taken was meant to be comprehensive in at least two dimensions:

i. Multiple, independent sources of information were sought (health care experts, the literature, and parents). It was felt to be important to include health care experts and parents in this early stage because they will be the primary stake holders in the final

questionnaire. That is, it will be the health care professionals who choose to use the instrument and the parents who choose to complete it. The instrument must, therefore, be relevant to them. One way of helping ensure this is by involving them in all steps of the research, including the generation of items. Furthermore, to ensure that the views of all types of parents were expressed, a broad representation of various children's ages and etiologies was sought.

ii. The topics covered in the interviews and focus groups were guided by the initial definition of the concept so as to be comprehensive about the coverage of all aspects of health. This was particularly useful in the focus group setting for which the parents would talk at length about a given area and might have otherwise neglected areas such as levels of independence or emotional health.

4.3.1 Limitations

One of the potential limitations of this process was that some areas of discussion may have been missed during the focus group discussions. For example, one of the known disadvantages of focus groups is that participants might be less likely to talk about particularly sensitive or personal topics. This possibility was minimized by having small groups involving individuals who share a somewhat common experience. As well, it would be unlikely that a particular topic would have been consistently ignored by five separate focus groups. Furthermore, it was the investigator's impression during these focus groups that the parents were eager to talk and that they all appeared very honest, comfortable and forthright in discussing very personal issues. Nonetheless, it is possible that certain aspects of health concerns might have been intentionally avoided in the discussions.

Another reason that certain health topics might have been ignored during the focus group sessions is the potential lack of representation of certain types of patients with unique health issues. This was recognized from the outset as a potential limitation and attempts were made to have as broad a representation of ages and etiologies as possible. However, it is possible that the categories used to ensure this broad representation were not adequate and that there exists an ignored group who would have provided added, unique items.

Another limitation of this process was in the transcribing phase. That is, the information from the focus groups was transcribed by the investigator using two methods. The first was the transcribing of the ideas mentioned by the parents during the focus group discussion itself. This was written on a blackboard in the conference room, to serve as a reference and reminder to the groups as to what had already been discussed. At the end of the session, the contents of this were transcribed onto paper as a permanent record. The second, and more thorough, method of transcription involved the investigator reviewing the videotape of each session and transcribing directly from what was said. The main limitation of both of these is the potential inaccuracy in transcribing in exact detail that which was said and intended by the parents. At least with the blackboard transcription the parents had the opportunity to correct what was written if it

did not adequately express their idea. This was not possible for the videotape transcription.

4.4 Summary

By seeking information from three separate sources (experts, the literature, and parents) a comprehensive list of 187 items was generated. The list was generated such that it would cover all relevant areas of health, recognizing that hydrocephalus affects children of all ages and of varying etiologies. Further steps in this project would involve reducing the list to a more practically feasible number of items for a final health status instrument.

5. Item Reduction

Various methods for reducing an item list have been reported in the literature and these have been broadly categorized as either *clinimetric* or *psychometric*.[56, 139-142] The technique that was used for this instrument borrowed from the clinimetric method used by Marx et al..[142] Briefly, this involved developing a severity-importance questionnaire (the item reduction questionnaire) to allow the respondent to rate both the severity and importance of each item. These were combined to create a severity-importance score which was then used to rank and select the most important items.

5.1 The Item Reduction Questionnaire

5.1.1 Construction of Item Reduction Questionnaire

5.1.1.1 Items

All of the 187 items generated during the previous stage of this project were formatted into one of two types of statements. The first format was a statement beginning with "My child…". For example, "My child has many friends" or "My child has a short attention span". There were 165 such items. The second format was a statement regarding parental concerns and began with "I am…". For example, "I am concerned my child will need his/her shunt to be lengthened" or "I am worried about my child's ability to live alone in the future". There were 22 such items.

A final list of questions was added at the end for demographic purposes and to help in item reduction. These asked how many shunt operations the child has had, the age of diagnosis with hydrocephalus, and who the primary care giver of the child was. Parents were also asked to provide a global rating of their child's health on 6-point scale with the following descriptors:

1- very severely impaired..... O
2- severely impaired..... O
3- moderately impaired..... O
4- mildly impaired..... O
5- very mildly impaired..... O
6- not at all impaired (normal health)..... O

This would later be used to ensure that sampling included children from all ranges of health status.

5.1.1.2 Response Options

For each of the 187 items, a 5-point adjectival scale was used for the respondent to assess how true the given item statement is (the "severity" of the item) and how important this is to their child's well-being (the "importance" of the item). The descriptors used were, in order:

Not at all
A little bit
Somewhat
Quite a bit
Very

The respondent would then check a circle that best corresponded to their feelings about the severity and importance of the item. Parents were asked, when responding, to consider how their child had been during the previous 4 weeks.

5.1.2 Pilot Testing of Item Reduction Questionnaire

A preliminary version of the questionnaire was administered to three parents of children with hydrocephalus, selected from the Hydrocephalus Database and known to live within the greater Toronto area. They were contacted by telephone by the investigator and all three agreed to participate.

The three parents, individually, came into the Hospital and were presented with a sealed envelope. This was meant to replicate the actual mailed packaged that would be received at home. This envelope contained: a cover letter (see

Appendix E – Cover Letter for Item Reduction Questionnaire), two copies of the item reduction questionnaire with instructions (one clearly labeled for "MOTHER" and one for "FATHER" – see **Appendix D - Item Reduction Questionnaire (selected items)**), a consent form, and a pre-addressed, stamped return envelope. The investigator simply instructed the parents to open the envelope and follow the directions. No other verbal instructions were given, although the parents had been made aware that the purpose of this research project was to investigate the health status of children with hydrocephalus. The investigator then left each parent alone in the room to complete the questionnaire. The investigator returned initially after 40 minutes, then at regular intervals afterwards if the parents had still not finished the questionnaire. The investigator's pager number was provided to the parents if they needed to contact him at any time.

Once the parent had finished, the investigator began an informal, semi-structured interview with the parent. The topics covered included the following:

- general esthetics of the entire package
- clarity of instructions and cover letter
- clarity and meaning of individual questions and the response options
- important items they felt were missing
- their general willingness to complete such a questionnaire if received at home

5.1.2.1 Results of Pilot Testing

The three parents completed the questionnaire in a mean of 41 minutes (38 – 45 minutes). They all felt the instructions and package were easy to understand and complete. They felt the time burden of the questionnaire, while long, was not sufficiently burdensome that it would have prevented them from completing this at home. They felt that most parents would welcome the opportunity to complete such a questionnaire.

There were, however, some suggestions for change. These involved the use of certain words that could be replaced by simpler, easier understood words. They also suggested being clearer in the instructions about what time frame they were to refer to when responding to the items. When asked about their preference for grouping of items into similar categories (i.e., having all physical health questions presented together), the opinions were mixed. One parent felt he would have preferred the grouping format, while the other two felt that having mixed grouping forced them to think more carefully about each question.

The changes recommended by the parents were implemented and this resulted in the creation of the final version of the item reduction questionnaire. (see **Appendix D - Item Reduction Questionnaire (selected items)**)

5.1.3 Administration of Item Reduction Questionnaire

5.1.3.1 Sample Selection

The final version of the item reduction questionnaire was administered as a mailed

survey to parents of children with hydrocephalus. The estimated required sample size was calculated based on the ability to produce a proportion estimate with a specific level of precision. Therefore, in order to produce a confidence interval of less than ± 0.15 around a proportion estimate, a minimum of 50 children would need to be recruited.

Patients, diagnosed with hydrocephalus for at least 6 months, were initially chosen by selecting names from the Hydrocephalus Database, excluding those who had already participated in earlier stages of this project. Twenty names were randomly selected from each of the following age groups: 3-4 years, 5-7 years, 8-12 years, 13-18 years. The list was reviewed to ensure adequate representation from each of the following categories of etiologies: congenital/aqueductal stenosis, myelomeningocele, and other. However, random selection was not stratified to these etiology categories.

All parents were called by the investigator prior to the mailing of the questionnaire. This served three purposes:

- to provide a personal introduction to the research and the investigator
- to ensure interest and willingness to participate
- to ensure accuracy of mailing information

Attempts were made to call all 80 parents who had been selected. Nine could not be reached, either because of an incorrect phone number or no answer despite multiple attempts. Of the 71 parents who were reached by telephone, one did not wish to participate and, in another case, the child had died two days prior. This left 69 of the 80 parents who were reached and willing to participate in the study (86.3%). In all of these cases, current, accurate mailing information was obtained.

5.1.3.2 Questionnaire Mailing and Follow-Up

All 69 parents were mailed a questionnaire package. This consisted of a large brown envelope, bearing the Hospital for Sick Children logo. Addresses were printed up on labels but regular, adhesive postage stamps were used (rather than business stamping). The envelope contained: a cover letter (see

Appendix E – Cover Letter for Item Reduction Questionnaire), two copies of the item reduction questionnaire with instructions (one clearly labeled for "MOTHER" and one for "FATHER" – see **Appendix D - Item Reduction Questionnaire (selected items)**), a consent form, and a pre-addressed, stamped return envelope. These were all held together by paperclip so that they would be removed from the envelope as a single unit. The cover letter was individually signed by the investigator in ink, along with a personal, hand-written note of thanks. The cover letter asked that both mother and father, if available, should complete separate copies of the questionnaire. The investigator's pager number was prominently displayed with instructions for the parents to call him at any time if they had any questions or concerns.

Between 7 to 10 days following the initial mailing, parents were called back to remind them about the questionnaire and to answer any questions they had. This also provided an opportunity to ensure that the parents had understood the instructions provided. If responses still were not received, parents were called back at repeated 7 to 10 day intervals. After 6 weeks, a letter was sent by overnight courier to those who had not yet responded. This was supplemented by another round of telephone calls. All calls were made directly by the investigator.

5.1.4 Survey Response

In total, 61 of the 69 sets of parents (88.4%) responded to the mailed survey. However, one set of parents returned a blank questionnaire since their child had very recently suffered a severe recurrence of a brain tumour and was in a moribund state. That left a total of 60 useable sets of responses for a final response rate of 87.0%. Of these, 36 sets returned questionnaires with both mother and father responses, for a total of 96 individually completed questionnaires.

The reasons for non-response among the eight who did not return their questionnaires were:

1..... mother was very poor in reading English and could not complete the questionnaire

1..... investigator was unable to reach the parents by telephone despite eleven attempts at various times of the day

2..... parents stated they were no longer interested in the project and would not be completing the questionnaire

4..... parents said questionnaire had been completed, but it was never received by the investigator

The characteristics of the 60 children for whom responses were received were as follows:

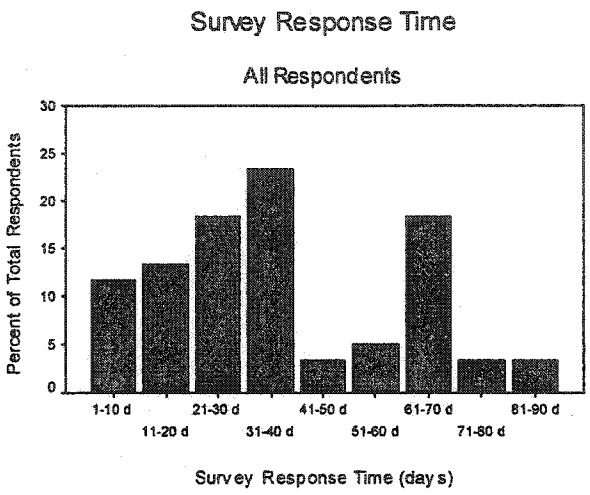
TABLE 5.1. Characteristics of survey respondents

	Mean ± SD	Min - Max
Age	8.3 ± 4.5 years	3 – 17
Age at diagnosis of hydrocephalus	14.2 ±29.2 months	birth – 14 years
Duration since hydrocephalus diagnosis	85.0 ± 52.8 months	6 months – 16 years
Number of shunt operations	4.4 ± 7.1	1 – 48
Primary care-giver global health rating	4.2 ± 1.5	1 – 6

5.1.4.1 Response Time

Of the 60 responses received, the mean time from mailing of the survey to receipt of the completed survey (*response time*) was 36.6 days (min-max = 8 – 85 days, median 31 days). The distribution of response time is demonstrated in the following graph:

FIGURE 5.1



5.1.4.2 Respondents' Ages and Etiologies

The 60 respondents represented the following age and etiology categories:

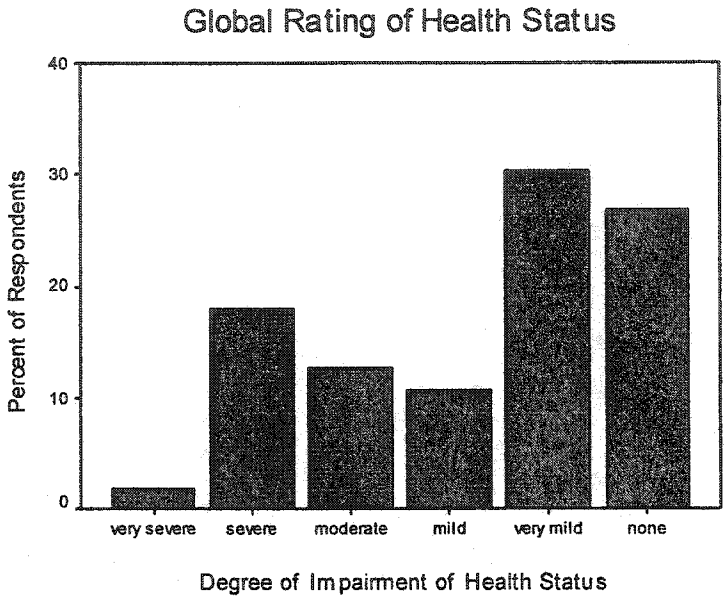
TABLE 5.2. Etiology and ages of survey respondents

Age Group	Etiologies			Total
	Congenital/ aqueductal stenosis	Myelo-meningocele	Other	
3-4 years	6	2	9	17
5-7 years	4	4	7	15
8-12 years	5	4	7	16
13-17 years	3	4	5	12
Total	18	14	28	60

5.1.4.3 Global Health Rating of Respondents

The responses to the global health rating questions demonstrated a reasonably broad spectrum of disease severity:

FIGURE 5.2



5.1.4.4 Single- versus Double-Parent Responders

Twenty-four responses were completed by only one parent (primary care-giver in all cases). In 36 responses, both mother and father completed separate copies of the questionnaire. The characteristics of single- versus double-parent responders are listed below:

TABLE 5.3. Single- versus double-parent responders

	Single-Parent Responders (mean ± SD)	Double-Parent Responders (mean ± SD)	p-value (independent sample t-test, 2-tailed)
Age (years)	7.5 ± 4.6	8.8 ± 4.5	0.28
Age at diagnosis (months)	16.6 ± 35.6	12.7 ± 24.7	0.63
Number of shunt operations	2.7 ± 2.2	5.4 ± 8.8	0.18
Global health rating	4.1 ± 1.6	4.3 ± 1.5	0.71

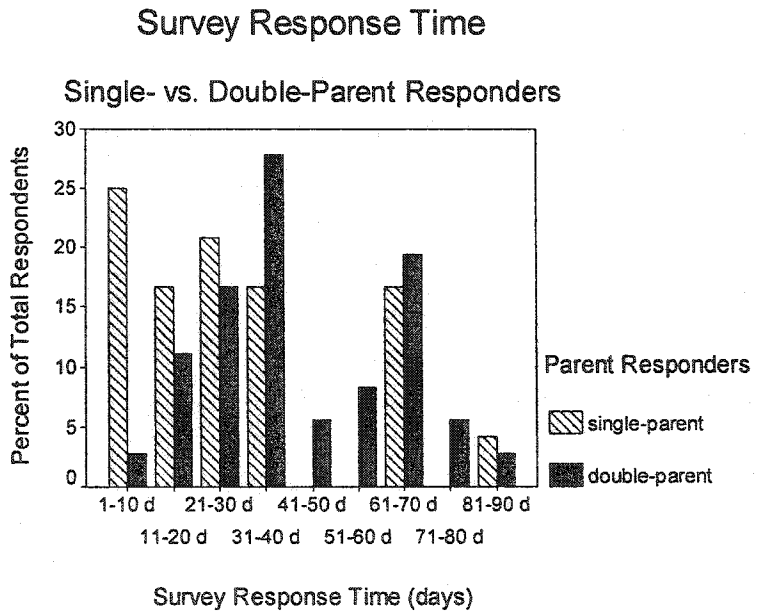
The distribution of etiologies is listed below:

TABLE 5.4 Etiologies for the responders

Type of Respondents	Etiologies			Total
	Congenital/ Aqueductal Stenosis	Myelo-meningocele	Other	
Single-parent	8	5	11	24
Double-parent	10	9	17	36
Total	18	14	28	60

The mean response time for the single-parent responders was 29.1 ± 21.4 days, while for the double-parent responders it was 41.6 ± 19.9 days (p=0.02, independent sample t-test, 2-tailed). The distribution of response times for the single- and double-parent responders is shown below:

FIGURE 5.3



The apparent delay in double-parent response time was likely due to the added time required waiting for an additional person to complete the questionnaire.

5.2 Selecting the Items

5.2.1 Method of Selecting Items

5.2.1.1 Assigning Domains

The first step in selecting the items was to assign each item to a unique domain. It was the investigator's opinion that, having spoken with parents at length during the item generation and item reduction phase, the initial concept of 9 domains could be reduced to five:

Physical Health
Social Health
Cognitive Health
Emotional Health
Parental Concerns

Therefore, each of the 187 items was assigned to one of these domains exclusively. This was done initially by the investigator, then verified by a content expert (a pediatric neurosurgeon). All further analysis was performed within the list of items for each domain. As well, of the 96 individually completed questionnaires that were available, only 60 were used for item selection analyses. That is, if both parents had completed questionnaires for the same child, only the response from the primary care-giver was used.

5.2.1.2 Eliminating Redundancy

Within each domain, all items were examined and those that appeared to tap very similar areas of content were reviewed closely for statistical redundancy. This followed a two step process. First, the Pearson correlation of the severity scores for the two items was checked. If this was greater than 0.6, then the Pearson correlation of the severity-importance (SI) scores was checked. The severity-importance (SI) scores were calculated by multiplying the severity score (ranging from 0 (representing better health status) to 4 (representing worse health status)) by the importance score (ranging from 0 (representing no importance) to 4 (representing very important)). Thus, the range in SI scores was 0 (not at all important or severe) to 16 (very important and severe). If the correlation of the SI scores was also greater than 0.6, then the items were considered redundant and one was selected for elimination. Given the theoretical concern about the effect of multiplication versus addition on the correlation of variables, the correlation between the SI scores was also calculated using SI scores obtained through simple addition (as opposed to multiplication). In every case, while the actual magnitude of the correlation of the SI scores changed slightly, the difference was very minimal and did not affect the selection criteria.

If the items were felt to be statistically redundant, then the following criteria were used to eliminate all but one of the items. First, a histogram of the distribution of the severity responses was created. Preference was given to items that had a more balanced spread of scores, across the range of the scale. Second, the item with a clearer and more generally applicable wording was given preference.

5.2.1.3 Selecting the Most Important Items

After the initial round of elimination, the mean SI scores (obtained through multiplication) for the remaining items were ranked within each domain. From this list, the most important items, i.e., the highest ranked items, were selected in order to represent a broad range of sampling content. The distribution of the severity responses was examined for each of the chosen items to ensure that no more than 70% of the respondents chose one of the extreme responses.[140]

5.2.1.4 Content Expert Review

The preliminary list of items selected from the above process was sent to six content experts. These were experienced pediatric neurosurgeons across North America (3 from Toronto, and 1 each from Salt Lake City, New York City, and Detroit) who have an established interest in clinical hydrocephalus research. They were asked their opinion of the items selected, specifically:

- the usefulness and ease of use of the instrument
- the appropriateness of items
- neglected areas of content
- the appropriateness of the categorization of items within their selected domains

Based on their recommendations, two individual items were added to the preliminary instrument: an item about headaches and about seizures. These items were on the initial list of 187-items, but had not been selected according to the previous selection algorithm.

5.2.2 Preliminary Results of Item Selection

Following the protocol described above, a total of 62 items were selected from the initial list of 187. The number of items within each domain was:

Physical Health.....15 items
Social Health..... 13 items
Cognitive Health.....12 items
Emotional Health.....13 items
Total (sum of the above 4 domains)..... 53 items
Parental Concerns.....9 items

The actual items themselves are listed in **Appendix F – 62-Selected Items for Reliability and Validity Testing.** Scoring for each of the items was done on a scale from 1 to 5, with 1 being "better health status" and 5 being "worse health status". Domain scores were than calculated by simple addition of all item scores within the domain. The Total score was calculated as the simple sum of the Physical, Social, Cognitive, and Emotional Health domains. Parental Concerns was not included in the total score, since it was felt to be a conceptually different domain.

5.3 Preliminary Measurement Properties

Using the severity scores already collected in the item reduction questionnaire, the following measurement properties were tested for the 62 items that were selected. These were, obviously, very preliminary. They did, however, provide some insight into the expected behaviour of the items that had been chosen. All analyses for this project were carried out using the SPSS 8.0 Advanced Statistics Package (SPSS Inc., Chicago, Illinois, USA).

5.3.1 Internal consistency

Internal consistency, using Cronbach's alpha, was assessed within each domain and for the instrument as a whole:

Physical Health.....0.91
Social Health..... 0.81
Cognitive Health.....0.90
Emotional Health.....0.79
Total (simple addition of the above 4 domains)..... 0.94
Parental Concerns.....0.86

5.3.2 Construct validity

A Pearson correlation of each domain score, calculated by simply adding all severity responses, with the parent's global rating of health (on the 6-point scale) was performed:

Physical Health.....0.85
Social Health..... 0.70
Cognitive Health.....0.55

Emotional Health.....0.50
Total (simple addition of the above 4 domains)..... 0.80
Parental Concerns.....0.36

5.3.3 Inter-rater reliability

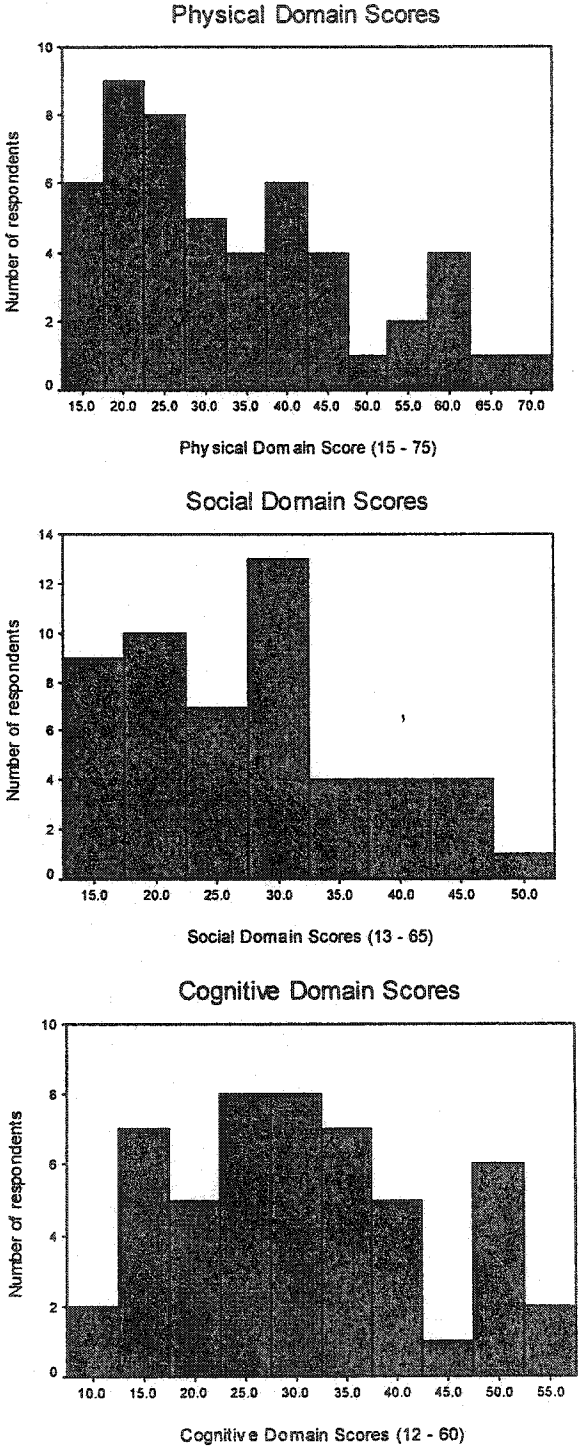
Among the double-parent responders, there were 19 mother-father pairs who responded completely to all items. Amongst these pairs, the ICC for domain scores between mother and father respondents was calculated. These are listed below, with their 95% confidence intervals:

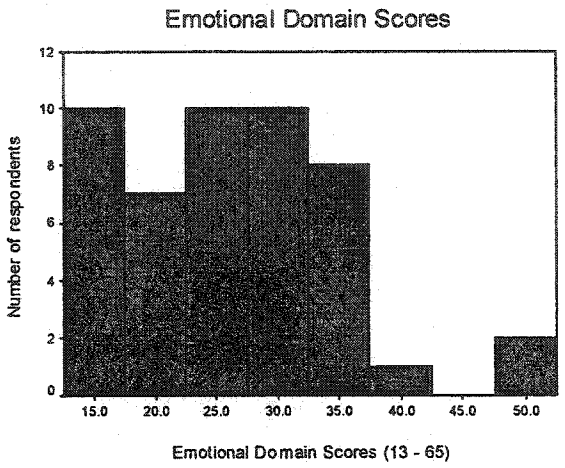
Physical Health.....	**0.91**	**(0.82 – 0.96)**
Social Health.....	**0.86**	**(0.73 – 0.93)**
Cognitive Health.....	**0.79**	**(0.58 – 0.90)**
Emotional Health.....	**0.74**	**(0.49 – 0.87)**
Total (simple addition of the above 4 domains).....	**0.89**	**(0.70 – 0.96)**
Parental Concerns.....	**0.78**	**(0.60 – 0.89)**

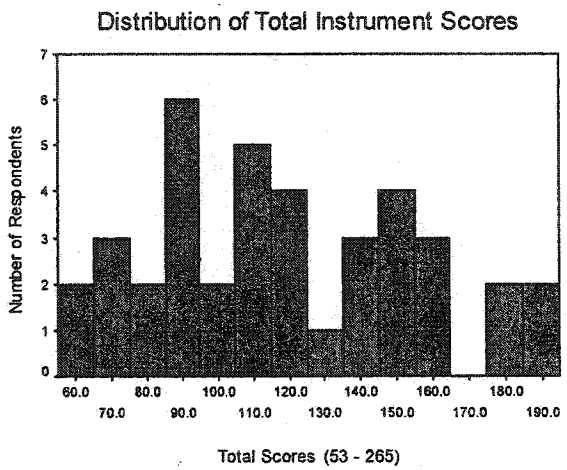
5.3.4 Distribution of scores

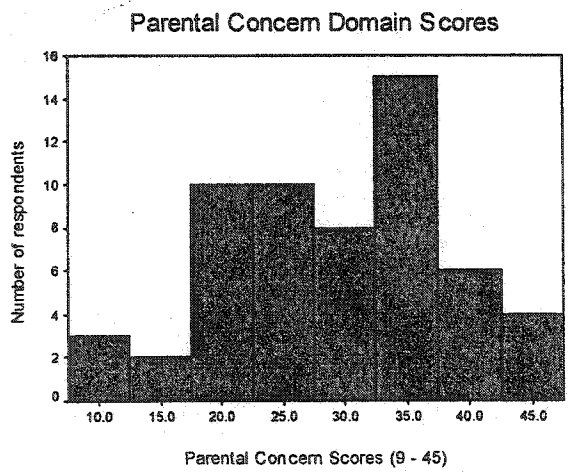
The distribution of domain scores and the total score for the instrument approximated a normal-like distribution with good spread of the data. The graphs below demonstrate this. Note that for each graph, the right side of the x-axis represents "worse health status":

FIGURE 5.4. Distribution of domain scores









5.3.5 Validity Across Groups

One of the foreseen difficulties with developing a hydrocephalus health status

measure is ensuring usefulness of the measure across the diverse age and etiology groups. The process of item reduction used responses from all these groups together. However, to provide some evidence of the usefulness of this instrument across these diverse groups, the correlation of the domain and total instrument scores with the parents' global health ratings was compared *within each of these groups*. These are listed below:

TABLE 5.5. Pearson Correlation of Domain Scores with Parent Global Rating: Age Groups

Health Instrument Domains	Age Groups			
	3-4 years (n=17)	5-7 years (n=15)	8-12 years (n=16)	13-17 years (n=12)
Physical	0.83	0.77	0.90	0.84
Social	0.67	0.74	0.82	0.71
Cognitive	0.58	0.32	0.58	0.48
Emotional	0.74	0.31	0.31	0.45
Total	0.69	0.83	0.93	0.62
Parental Concerns	0.24	0.15	0.19	0.71

TABLE 5.6. Pearson Correlation of Domain Scores with Parent Global Rating: Etiology Groups

Health Instrument Domains	Etiology Groups		
	Congenital/ Aqueductal Stenosis (n=18)	Myelomeningocele (n=14)	Other (n=28)
Physical	0.84	0.70	0.83
Social	0.85	0.36	0.80
Cognitive	0.56	0.25	0.55
Emotional	0.57	0.16	0.49
Total	0.81	0.46	0.83
Parental Concerns	0.13	0.36	0.57

These data indicate that the selected items appear to maintain a moderate-to-strong degree of correlation within each of the age and etiology groups. This suggests preliminary evidence of construct validity, i.e., that these items do appear to be measuring health status within each group. The only group of concern was the myelomeningocele group, in which all correlations, except for physical, were substantially lower. Part of this may have been explained by one outlying measure. This was a case in which a child is wheelchair-bound and was given a global health rating of only "mildly" impaired.

However, when the analysis was repeated with this case excluded, the results were very similar. Alternatively, therefore, it is possible that this might represent a genuinely different way in which parents of children with myelomeningocele rate their children's health.

5.4 Pilot Testing of Questionnaire

The 62 selected items were formatted into a two-page questionnaire with instructions for completion. This questionnaire is hereafter referred to as the Hydrocephalus Outcome Questionnaire (HOQ). In order to test out the HOQ, parents of children with hydrocephalus were identified through the Hydrocephalus Database. Parents were identified according to the previously used criteria, with the added criteria that they must live within the greater Toronto area. Parents were initially telephoned by the investigator and informed about the research project and its purpose. They were asked if they would be interested in coming to the Hospital for a 1-hour research session.

Those parents who agreed to participate met with the investigator on an individual basis in a conference room at the Hospital. They were told, once again, about the research project and what was needed of them during the research session. The parents were then given a copy of the questionnaire and asked to complete it. A brief set of verbal instructions, lasting about 1 minute, was given to the parents. The investigator then left the room to allow the parents to complete the questionnaire in private. Although the parents were not aware of it, they were timed to see how long it took to complete the questionnaire.

Upon completion of the questionnaire, an informal, semi-structured interview was conducted by the investigator. The topics covered included the following:

- general esthetics of the questionnaire
- clarity of instructions
- clarity and meaning of individual questions and the response options
- important items they felt were missing

Parents were encouraged to be critical of the questionnaire. It was stressed by the investigator many times that the purpose of this session was to get constructive feedback and to make changes to questionnaire. This was repeated to make the parents more comfortable in expressing any criticism of the questionnaire.

The parents were also asked to complete some other health status questionnaires in anticipation of their use in later validity testing. These were the Health Utilities Index – 2 (HUI-2),[90] the Functional Independence Measure for Children (WeeFIM),[143] the Impact-on-Family Scale (IFS),[144] and the Strengths & Difficulties Questionnaire (SDQ),[145]. The completion of the HUI-2 and the WeeFIM required assistance, while the IFS and SDQ were completed by the parents unassisted. Instructions were given to the parents about how to complete each of these questionnaires and they were timed as to how long each questionnaire took to complete. For a more complete discussion about these instruments, please see the **Chapter 7**. The parents were asked at the end of the

session the following:

- Which of the 5 questionnaires was the easiest to complete?
- Which of the 5 questionnaires asked questions that are most important to your child's health?

Finally, parents were asked to provide a global rating of their child's overall health, physical health, social health, cognitive health, and emotional health. These were all asked on separate 6-point response scales ranging from "very severely impaired" to "not at all impaired (normal health)".

At the conclusion of the session, parents were asked if they would agree to complete some of these questionnaires (those that did not require assistance) at home in approximately two weeks time. If they agreed, their mailing address was confirmed and recorded. At the conclusion of the session, parents were given a Hospital parking pass as a token of appreciation for their time.

5.4.1 Results of Pilot Testing

A total of 6 sets of parents (7 individuals) attended the initial pilot testing sessions. The mean duration of time required to complete the HOQ was 12 minutes (min-max = 10 – 17 minutes). All participants found the questionnaire clear and did not find any difficulty in its use. Although the participants were repeatedly asked about suggestions for improvement, no substantive suggestions were made during the semi-structured interviews. The mean total duration of the sessions was 40 min (min-max = 34-47 minutes).

Preliminary results of test-retest reliability were calculated using the repeat data from 7 individual parents. They repeated the questionnaire a mean of 16.6 days apart (min-max = 10-28 days). The intraclass correlation coefficients (ICC) (95% CI) for each of the domain and total scores were:

Physical Health.....**0.92** **(0.61-0.99)**
Social Health..... **0.88** **(0.40-0.98)**
Cognitive Health.....**0.90** **(0.72-0.99)**
Emotional Health.....**0.38** **(0.00-0.86)**
Total (simple addition of the above 4 domains)..... **0.86** **(0.40-0.97)**
Parental Concerns.....**0.94** **(0.72-0.99)**

In addition, 4 mother-father pairs completed the questionnaire at the same time, allowing for a preliminary comparison of inter-rater reliability. The ICC's were:

Physical Health.....0.79 (0.00-0.98)
Social Health..... 0.85 (0.13-0.99)
Cognitive Health.....0.40 (0.00-0.97)
Emotional Health.....0.61 (0.00-0.96)
Total (simple addition of the above 4 domains)..... 0.73 (0.00-0.99)
Parental Concerns.....0.87 (0.00-0.99)

Because of the very small sample size, all the 95% confidence intervals were exceptionally large making interpretation of these numbers difficult.

The other questionnaires (the HUI-2, the IFS, the WeeFIM, and the SDQ) were generally completed in less time (most under 5 minutes each). Three of seven parents felt that the HOQ was the easiest to complete while six of seven felt that it asked questions that were most important for their child's health.

Test-retest reliability was able to be calculated on the global ratings, the IFS, and the SDQ:

Global Ratings of Health:
Overall health..... 0.83 (0.30-0.97)
Physical health..... 0.62 (0.00-0.92)
Social health..... 0.59 (0.00-0.92)
Cognitive health..... 0.97 (0.86-1.00)
Emotional health..... 0.53 (0.00-0.90)

IFS Total Score..... 0.73 (0.12-0.95)

SDQ Total Score..... 0.73 (0.11-0.95)

5.5 Discussion

There is no consensus on the best way to select items for a health status instrument.[56, 139-142] The clinimetric approach attempts to select out the items deemed most important by the respondents. The psychometric approach attempts to eliminate items based on factor analysis structure and the distribution of endorsement. The process used here was guided primarily by parental responses indicating which items they felt were most important and affected their children the most. This was clearly a strategy borrowed from the clinimetric approach. However, elements of the psychometric approach were also utilized. For example, in deciding which of two similar items to retain, the response distribution was examined to select the one with better spread and to eliminate those with greater than 70% endorsement of one of the extreme responses. This

combination of approaches was an acknowledgement that neither approach can be considered in isolation and both have their advantages.

Beyond the controversy of the different approaches, there is also little consensus within each of these approaches as to how to optimally proceed with item reduction. For example, using the clinimetric method of severity-importance questionnaires, it is not clear what severity options should be presented to respondents, i.e., a dichotomous response (as described by Juniper et al.)[56] or a multi-category response option (as described by Marx et al. and used here).[141] As well, it is not clear what would be the best way to mathematically combine the severity and importance responses, e.g., addition or multiplication. This latter controversy was investigated empirically by Marx et al. who did not find a substantial difference in the final selected items based on the method of combining the scores.[141, 142] The current study demonstrated similar results, with the use of addition rather than multiplication having little impact on the ranking of the SI scores and the selection of items.

The 62-items which were ultimately selected were tested for preliminary evidence of their psychometric properties. This was used purely as an exploratory analysis, given the limitation in sample size. This analysis did suggest, however, reasonable evidence of internal consistency and inter-rater reliability, although the confidence intervals for some of the estimates of the latter were quite wide. An attempt was also made to look at evidence of construct validity by comparing the severity scores to the general health ratings. However, the results were not very compelling because of the reliance on the global health ratings, which are likely not the best means of assessing health status in this population. Therefore, although some of these correlations were in the moderate-to-strong range, little confidence can be put into the interpretation of this.

Regardless of the limitations of this preliminary testing of the instrument's psychometric properties, one consistent and troubling feature was the relatively low reliability (by all measures of reliability) of the Emotional Health domain. In later sections of this work it will become clear that this was a legitimate finding and, in retrospect, perhaps more should have been done at this stage of the research to remedy it.

5.5.1 Limitations

One of the limitations of the item reduction process is the inherent lack of consensus in the optimal approach to use (see previous discussion). Based on the available literature and empirical experience, this appears to be an unavoidable limitation.

The item reduction questionnaires were administered through a mail survey. While this was the most efficient way to reach a large enough sample size, it has some potential disadvantage. Other than the obvious lack of 100% response rate, it can not be guaranteed that the parents completely understood the instructions or what exactly was intended of them when completing the questionnaire. The task involved was somewhat complex, requiring them to make two separate judgments about a statement and then record these responses on separate 5-point scales. As well, the task was long, requiring roughly three-quarters of an hour of dedicated time. By completing the survey at home,

the parents would not have the opportunity to immediately address any concerns they might have had about the questionnaire in general or specific items in particular. In order to try to address these limitations, the following procedures were used:

i. The questionnaire was pilot tested in a way that simulated the mail survey situation as closely as possible. This gave the investigator some confidence that there should be no major areas of ambiguity or difficulty with the questionnaire.

ii. The investigator made at least two separate telephone communications with the parents (before and then after the questionnaire was sent). This provided a planned and consistent opportunity for parents to ask questions and raise concerns. As well, it was made clear in the cover letter and instructions that the investigator was available by pager 24 hours a day to answer questions.

Despite these procedures, the parents were still ultimately left alone at home with the questionnaire. Even if they did completely and correctly understand the questionnaire, it is difficult to know what effect daily home distractions might have had on their ability to concentrate on this fairly lengthy and moderately complex questionnaire.

5.6 Summary

By surveying 60 sets of parents with a severity-importance item reduction questionnaire, the initial list of 187 items was reduced to a 62 item instrument called the Hydrocephalus Outcome Questionnaire (HOQ). These items represent four separate health status domains and the additional domain of Parental Concerns. Further work in this project would involve subjecting the HOQ to tests of its psychometric properties (reliability and validity).

6. Instrument Reliability

Reliability is the reproducibility of a measure and can be assessed in different ways.[84] It is also a measure of how well an instrument is able to differentiate among subjects while displaying small measurement error. In this sense, the reliability of an instrument can be very dependent on characteristics of the sample to which it is being administered. Reliability is the minimum requirement for a useful health status instrument. In this chapter the process of establishing the reliability of the HOQ is presented.

6.1 Methods for Reliability Testing

To examine the reliability of the HOQ, parents of eligible children were administered the questionnaire during routine clinic visits to the Hospital for Sick Children. Mailed repeat questionnaires were sent to the parents for completion approximately 2 weeks following the initial questionnaires.

6.1.1 Sample Selection and Contact

The parents of eligible children were identified by reviewing the list of clinic patients and cross-referencing this list with the Hydrocephalus Database. Approximately 1 week prior to the scheduled clinic visits, parents were telephoned at home by the investigator and informed about the research project. They were asked if they wanted to participate in the study, after being told what such participation would require from them. They were also told that, in exchange for their participation, they would receive a voucher for free parking at the Hospital. All clinic visits were routine and scheduled 6 – 12 months in advance, so the children were not being seen for any change in health status. The occasional patient whose clinic visit was scheduled as a result of a specific health concern was excluded from this study.

The minimum required sample size for determining the reliability was estimated to be 50 subjects. This was based on two separate methods of estimation. One method was based on the desire to produce a reliability coefficient with confidence intervals of less than $\pm$ 0.10.[50] Assuming a reliability coefficient of at least 0.80, then the required sample size would be 50 subjects. The second method was based on testing the null hypothesis of H_o: $\rho=\rho_o$ against the alternative hypothesis H_1: $\rho>\rho_o$ (where ρ_o is the minimally acceptable reliability coefficient).[146, 147] Under the assumptions of $\rho_o=0.70$, alpha = 0.05, beta = 0.20, and an expected reliability of 0.85, then the model predicts a required sample size of 43 subjects. Therefore, taking both methods into account, a minimum of 50 subjects were felt to be necessary for this reliability study.

6.1.2 First Administration of Questionnaire (Time 1)

Parents who agreed to participate met with the investigator or research assistant following their visit with the attending surgeon. The parents and the child were told about the study and its purpose and what would be required of them during the session. The parents were given consent forms and the children were given assent forms (if under 16 years old) or consent forms (if 16 years or older). After reading these forms, parents were asked again if they had any questions about the project and then were asked to sign the forms.

The questionnaire administration began with the HOQ. If both parents were present, they were given separate copies and asked to complete the questionnaires separately. Instructions were given in a standardized fashion and emphasized the following information:

- the format of the items
- the format of the response options
- the 4 week time interval that should be applied when responding to items

After completion of the HOQ, parents were then asked to complete several other questionnaires as part of the Validity testing (see **Chapter 7**).

The subject child was offered the opportunity to complete a child version of the HOQ if the following conditions were met:

- child was 12 years of age or older
- child was felt, by the parents, to have the cognitive capacity to complete the HOQ
- the parents offered their consent
- the child offered his/her consent

The child version of the HOQ was similar to the parent version except that the context of the items was changed from third-person (e.g., "My child has poor vision") to first-person ("I have poor vision"). Otherwise, the items and response options were identical. The children were given separate instructions about completing the questionnaire.

6.1.3 Second Administration of Questionnaire (Time 2)

Ten days following the completion of the first questionnaire, the parents were mailed a second identical questionnaire. All parents were told of this at the first administration and they were informed of the rationale for this repeat questionnaire. It was also stressed during the first administration that the second questionnaires should be done promptly (within one day of receipt) and that both parents should complete it separately.

The mailing package consisted of the following, all sent in a large-format Hospital for Sick Children envelope with the mailing address hand-written and stick-on postage stamps applied:

- a cover letter
    - explaining the purpose of the second administration
    - providing instructions to complete within one day
    - providing instructions for parents to complete it separately
    - pager number to contact the investigator with any questions or concerns
    - a hand-written "thank-you note" to the parents for their participation
    - personally signed by the investigator

- copies of the questionnaires, clearly labeled "MOTHER" and "FATHER" (and one labeled for the child, if appropriate)
- a stamped, pre-addressed return envelope

6.1.4 Follow-Up

Parents were telephoned by the investigator approximately one week after the mailing. This was to ensure that the package had arrived, to remind them to complete it (if not already done) and to answer any questions they had.

If packages were not received back at the Hospital within two weeks, another telephone call was made and this was repeated at weekly intervals. After four weeks, a repeat package was sent with a personalized letter asking for the parents to complete the questionnaires as soon as possible.

6.1.5 Analysis

The data were used to provide three estimates of reliability: a test-retest (TRT) reliability estimate, an inter-rate reliability estimate, and an internal consistency estimate.

The TRT analysis was performed using a repeated-measure analysis of variance (ANOVA). This provided the mean square values associated with the between-subject variance and the within-subject variance. From this, one could calculate the between-subject variance and the variance associated with the different times and the remaining residual error. Then, an intraclass correlation coefficient (ICC) was calculated.[148] This ICC included time as a random factor, so that its variance was included as part of the error term. Although time is sometimes left out of the error term in the denominator,[50] it was included here. The study design included two separate observations in time and these were treated as one would treat any two random observations, recognizing that the reliability coefficient would need to reflect the fact that in actual use, only one, randomly-selected observation time would be used.

$$ICC = \frac{\sigma^2_{subjects}}{\sigma^2_{subjects} + \sigma^2_{time} + \sigma^2_{error}}$$

This ICC was then taken as the reliability coefficient. This was repeated for the sub-score of each domain and for the total instrument score. As well, a standard error of measurement (StErMe – see **Section 2.4.1.2.7**) was estimated as the square-root of the mean-square error term from the ANOVA table.

An inter-rater reliability estimate was calculated in which each parent served as the repeated-measure for each child. The analysis used an ICC and a repeated-measures ANOVA, as described above. Each set of parents was regarded as being a selection from a larger population of parents and was treated, therefore, as a random factor. Following the example of Shrout and Fleiss in describing the different scenarios in which one can use ICC as an assessment of rater reliability, this methodology was most consistent with their "Case 1": "each target is rated by a different set of k judges, randomly selected from a larger population of judges."[149]

To calculate the internal consistency of the measure, Cronbach's alpha was used. Once again this was analyzed within each domain and for the instrument as a whole.

Finally, a generalizability study (G study) analysis was carried out.[50] This examined both Time (i.e., Time 1 and Time 2) and Parent respondent (i.e, mother and father) as within-subject sources of error. This was felt to be the most efficient way to examine both sources of error and their respective effects on the questionnaire responses. This was performed using a repeated-measures, two-way within-subject ANOVA. The two within-subject factors were *time* and *parent*, each with two levels. Each factor and an interaction term were analyzed for statistically significant effect using the Wilks lambda multivariate statistic at a significance level of 0.01 (to partially account for multiple comparisons).[150] Furthermore, the relative variance contribution of each component (Child, Parent, Time, and all interaction terms) was calculated. The percentage contribution of Child to overall variance can also be interpreted as a generalizability coefficient using absolute error (instead of relative error) in the denominator.[54] This has also been termed the dependability index.[55] A signal-noise ratio (Child variance component divided by the absolute error variance) was also calculated.[54] All analyses were performed for the total score and each domain sub-score.

6.1.6 Questionnaire Scoring

The initial method of scoring the HOQ involved simply summating the responses of each item from the 5-point scale, with each item receiving a score from 0 to 4. In this summation, responses indicating "worse health status" were given a value of 4 and those indicating a "better health status" were given a value of 0. Therefore, for some items which asked about favourable aspects of health (e.g., My child has many friends) a reverse coding was used, compared to those items which asked about negative aspects of health. This method of scoring provided a rather inconvenient metric for the HOQ Total Score (potential range of scores = 0 to 212) and for each of the domain scores (which ranged from 0 to as high as 60). As well, this scoring was rather counter-intuitive in that better health status was represented by lower numbers.

Therefore, in order to simplify the metric of the HOQ scores, the initial summated HOQ values were converted from their raw scores into a percentage score. This was achieved by dividing the raw summated score by the highest potential score available and then taking the complement of this. The values for any score (Total or domain scores) would now range from 0.0 (worse health status) to 1.0 (better health status).

6.1.7 Comparison to Severity-Importance Questionnaire

The methods used for Item Reduction (see **Chapter 5**) partly involved using the results of the severity-importance (SI) questionnaire to provide some preliminary evidence of the internal consistency and inter-rater reliability of the selected items. Therefore, as a matter of academic interest, a comparison of these early results to the current, more formal reliability testing results was made. It was hoped that, if some consistency was demonstrated, then the use of this criteria in item selection will have been justified, albeit retrospectively.

6.2 Results of Reliability Testing

6.2.1 Questionnaire Response

The parents of 94 eligible children completed the first questionnaire administration. Based on the comments of the parents during the completion of the questionnaires, it was decided to eliminate the 4 children who were under 5 years of age. These younger children had initially been included as an exploratory venture, recognizing that the HOQ was mostly likely only relevant to children 5 years and older.(see **Section 3.2.1.2**) As expected, the parents of these younger children had difficulty in completing the questionnaire because many questions were felt to be inappropriate to their child's age. Therefore, all further analyses and discussion are based on the sample of 90 children aged 5 years and over. The characteristics of these children are listed below:

TABLE 6.1: Characteristics of study participants

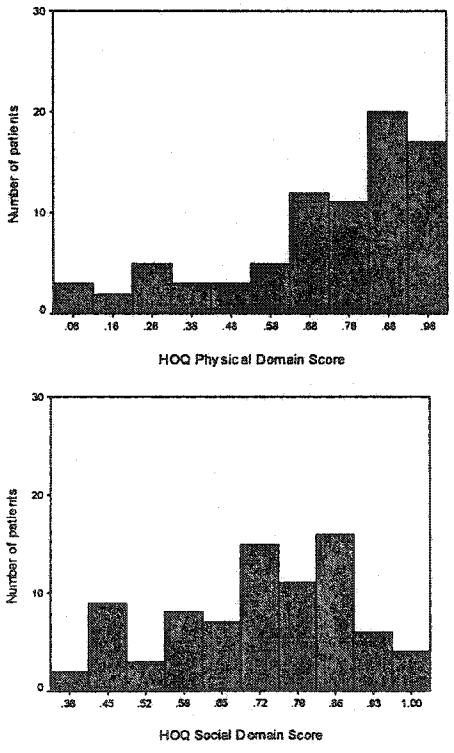
	Mean ± SD	Min - Max
Age	10.0 ± 3.5 years	5 – 18 years
Age at diagnosis of hydrocephalus	16 ± 35 months	0 – 36 months
Duration since hydrocephalus diagnosis	8.7 ± 3.7 years	1- 17 years

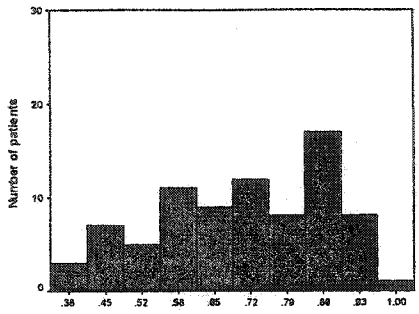
TABLE 6.2. Ages and etiologies for study participants

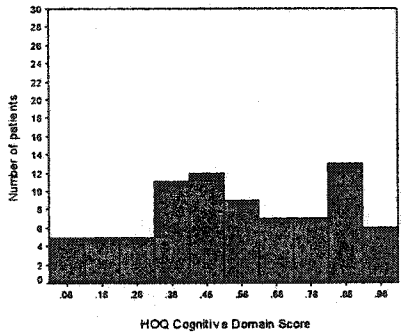
Age Group	Etiologies			Total
	Congenital/ aqueductal stenosis	Myelo-meningocele	Other	
5-7 years	8	3	13	24
8-12 years	14	8	20	42
13-17 years	10	1	13	24
Total	32	12	46	90

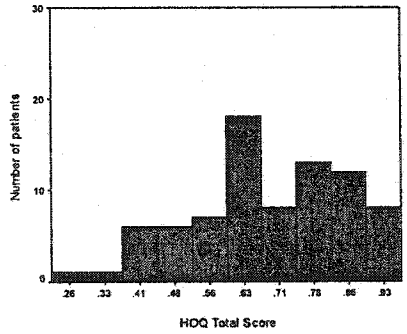
The distribution of the individual domain scores and the Total Score of the HOQ demonstrated good variability and approached a normal-like distribution, suggesting that an appropriate breadth of disease-severity was obtained (assuming, for the moment, questionnaire validity):

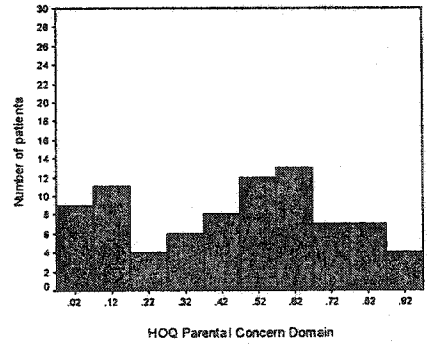
FIGURE 6.1. Distribution of domain scores











Of the 90 children, the parents (mothers and/or fathers) of 59 (66%) also completed the second questionnaire administration. However, of these 10 did not completely answer all questions and were, therefore, not usable. This left 49 (54%) with

completed and usable second questionnaires. This was despite an aggressive approach to data collection, as outlined earlier. Reasons for lack of compliance with the second set of questionnaires varied (10 parents unable to be reached, 2 parents refused, 5 parents claimed that they had completed and mailed the second questionnaire although it did not reach the investigator). The remaining stated that they would complete the second questionnaire, although it was never received by the investigator.

The characteristics of the 49 complete responders to the second questionnaire compared to non-responders is listed below. These did not seem to differ significantly:

TABLE 6.3. Characteristics of responders versus non-responders

	Responders	Non-Responders	p-value (2-sided t-test)
Age, years (mean ± SD)	10.4 ± 3.7	9.6 ± 3.3	0.2
Age at diagnosis of hydrocephalus, months (mean ± SD)	18 ± 42	14 ± 26	0.6
Duration since hydrocephalus diagnosis, years (mean ± SD)	9.0 ± 3.7	8.3 ± 3.7	0.4
HOQ Total Score, Mom, Time 1 (mean ± SD)	0.70 ± 0.16	0.65 ± .17	0.2

The distribution of ages and etiologies of the 49 responders also displayed appropriate breadth:

TABLE 6.4. Ages and etiologies of responders

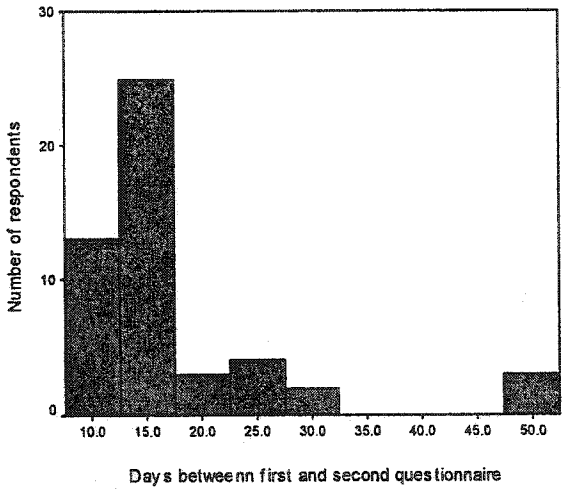
Age Group	Etiologies			Total
	Congenital/ aqueductal stenosis	Myelo-meningocele	Other	
5-7 years	3	2	7	12
8-12 years	6	5	12	23
13-17 years	7	1	6	14
Total	16	8	25	49

6.2.2 Reliability Estimates

6.2.2.1 Test-Retest

The test-retest ICC was calculated for all responders comparing the scores and subscores from the first and second questionnaire administrations. Amongst the 46 mothers and 30 fathers who completed both questionnaires, the mean duration between the completion of the first and second questionnaires was 17.6 days (min-max = 9 – 51 days). The distribution of the response time is shown below:

FIGURE 6.2. Distribution of response times



The ICC and 95% CI for each of the domain and total scores were:

	Mothers (n=46)		Fathers (n=30)	
Physical Health	**0.98**	**(0.97-0.99)**	**0.98**	**(0.96-0.99)**
Social Health	**0.85**	**(0.75-0.92)**	**0.86**	**(0.73-0.93)**
Cognitive Health	**0.92**	**(0.86-0.95)**	**0.93**	**(0.85-0.97)**
Emotional Health	**0.73**	**(0.56-0.85)**	**0.80**	**(0.63-0.90)**
Total	**0.93**	**(0.88-0.96)**	**0.96**	**(0.91-0.98)**
Parental Concerns	**0.86**	**(0.77-0.92)**	**0.91**	**(0.82-0.96)**

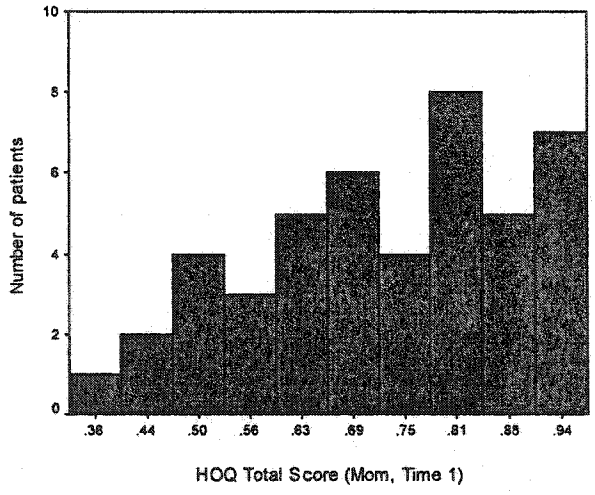
The standard errors of measurement were as follows:

	Mothers (n=46)	Fathers (n=30)
Physical Health.....	**0.03**	**0.03**
Social Health.....	**0.07**	**0.07**
Cognitive Health.....	**0.08**	**0.07**
Emotional Health.....	**0.07**	**0.06**
Total.....	**0.04**	**0.04**
Parental Concerns.....	**0.10**	**0.09**

6.2.2.2 Inter-rater Reliability

As another indicator of reliability, an inter-rater ICC coefficient was calculated between 45 pairs of mother-father respondents. This was obviously a biased sample, representing only those children who were living with both parents. The distribution of the HOQ Total Score for this subset of children is displayed graphically below:

FIGURE 6.3. Distribution of HOQ scores



This distribution of scores is perhaps slightly more skewed toward a better health status compared with the overall sample. The inter-rater reliability ICC's at Time 1 were:

Physical Health.....	**0.96**	**(0.94-0.98)**
Social Health.....	**0.84**	**(0.73-0.90)**
Cognitive Health.....	**0.76**	**(0.61-0.85)**
Emotional Health.....	**0.54**	**(0.31-0.70)**
Total Score.....	**0.88**	**(0.79-0.93)**
Parental Concerns.....	**0.67**	**(0.49-0.79)**

6.2.2.3 Internal Consistency

The Cronbach's alpha estimates of internal consistency for the instrument, at Time 1 for mother and father respondents were:

	Mothers (n=80)	Fathers (n=52)
Physical Health	0.93	0.93
Social Health	0.74	0.83
Cognitive Health	0.91	0.95
Emotional Health	0.73	0.80
Total	0.94	0.96
Parental Concerns	0.89	0.90

6.2.2.4 Generalizability Study

A G-study was carried out to examine the impact of Time and Parent on variability in HOQ responses. This was based on the sample of 26 patients for whom there were completed questionnaires by both parents at both times. The Wilks' lambda p-value for each of the factors within each of the domain scores and total scores are listed below (asterisks represent significant values <0.01):

	Time	Parent	Interaction (Time*Parent)
Physical Health	0.88	0.11	0.39
Social Health	0.74	0.23	0.05
Cognitive Health	0.16	0.53	0.40
Emotional Health	0.01*	0.94	0.03
Total	0.09	0.53	0.26
Parental Concerns	0.83	0.03	0.21

From this G-study, the effect of Time was found to have a significant impact on the HOQ Emotional Health responses. However, the following other effects approached significance: Time-Parent interaction for Emotional Health; Time-Parent interaction only for Social Health; and Parent only for Parental Concerns.

To better quantify the relative effects of various components, relative variance components were calculated, by determining the percentage contribution of each term to the total variance. These results, along with the signal-noise ratio, are presented in the following table:

TABLE 6.5 Variance components and signal-noise ratios

HOQ Domain	Child	Parent	Time	Child x Parent	Child x Time	Time x Parent	Child x Time x Parent	Signal -noise ratio
Physical	96	1	1	2	1	1	1	26.9
Social	82	1	1	1	5	1	10	4.4
Cognitive	72	1	1	16	2	1	9	2.6
Emotional	37	3	2	30	6	3	18	0.6
Total	88	1	1	4	2	1	5	7.6
Parental Concerns	70	3	1	18	1	1	8	2.3

Aside from Emotional Health, the variance of all other domains was very largely accounted for by Child variance, as would be expected. This was reflected by the high signal-noise ratios. For Emotional Health, however, a large part of the variance was due to Child-Parent interaction effect and the signal-noise ratio was very low (0.6).

6.2.3 Improving Instrument Reliability

The initial results of reliability testing were promising except for the Emotional Health domain subscore. It demonstrated relatively poor internal consistency and test-retest reliability. In order to improve this, the two items that had the worst impact on internal consistency ("My child copes well with his/her disability" and "My child feels confident") were removed. With this, the internal consistency improved to 0.76 and the test-retest ICC to 0.75 (0.57 – 0.85). This was only a modest improvement. A further step that was taken was the decision to combine the Social and Emotional Health domains into a single domain with a single subscore. This effectively increased the number of items in the domain to 24 and this also served to greatly improve the reliability estimates. The internal consistency of the Social-Emotional domain was 0.82 , the test-retest ICC was 0.84 (0.73 – 0.91), and the mother-father inter-rater ICC was 0.76 (0.61-0.85). This was considered a much more acceptable level of reliability. As well, by repeating the G-study for this new Social-Emotional domain, the following Wilks' lambda p-values were found:

	Time	Parent	Interaction (Time*Parent)
Soc-Emotional Health.....	**0.05**	**0.15**	**0.06**

As well, the following was the relative variance contributions for the new Social-Emotional domain, along with its signal-noise ratio:

TABLE 6.6. Variance components and signal-noise ratio

HOQ Domain	Child	Parent	Time	Child x Parent	Child x Time	Time x Parent	Child x Time x Parent	Signal -noise ratio
Social-Emotional	77	1	1	4	5	1	13	3.3

These values represented improvement compared to the earlier domains.

Despite this overall improvement in reliability, it should be noted that this change in domains was made only after considering its theoretical implications to the measurement construct. Specifically, as will be discussed in the Validity section, the domains of Social and Emotional Health were considered difficult to separate as distinct constructs. Therefore, aside from its impact on reliability, this change to a combined Social-Emotional domain also made sense theoretically. This notion was tested further in the Validity section and this will be discussed there (see **Chapter 7**).

The removal of the two items just mentioned also had a minor impact on the overall HOQ Total Score, which now consisted of 51-items. However, the internal consistency and test-retest reliability of the total score changed only negligibly (0.94 and 0.94 (0.89 – 0.97), respectively).

6.2.4 Reliability of the Child Version of the HOQ

There were 19 children aged 12 years and older who met the criteria for completion of the child version of the HOQ and who completed the first administration. Of these, 13 also completed the second mailed questionnaire. The reliability estimates for this sample were:

	Test-Retest ICC		Cronbach's alpha
Physical Health	**0.97**	**(0.90-0.99)**	**0.82**
Social-Emotional Health	**0.60**	**(0.05-0.87)**	**0.86**
Cognitive Health	**0.87**	**(0.62-0.96)**	**0.83**
Total	**0.84**	**(0.53-0.95)**	**0.90**
Personal Concerns	**0.85**	**(0.57-0.82)**	**0.94**

The scores obtained from the children were compared to their respective mothers' scores with ICC coefficients (inter-rater reliability) and Pearson correlations:

	Mother-Child ICC		Mother-Child Pearson correlation
Physical Health	**0.88**	**(0.71-0.96)**	**0.90**
Social-Emotional Health	**0.53**	**(0.10-0.80)**	**0.56**
Cognitive Health	**0.43**	**(0.00-0.74)**	**0.52**
Total	**0.63**	**(0.18-0.85)**	**0.71**
Personal Concerns	**0.08**	**(0.00-0.53)**	**0.08**

6.2.5 Comparison to Results of Severity-Importance Questionnaire

The following table shows a side-by-side comparison of reliability estimates obtained from the SI item reduction questionnaire and the HOQ reliability testing for internal consistency (Cronbach's alpha) and mother-father inter-rater reliability (ICC with 95% confidence intervals):

TABLE 6.7. Comparison of reliability estimates

Internal Consistency:	**SI Questionnaire**	**HOQ Reliability Testing**
Physical Health	**0.91**	**0.93**
Social Health	**0.81**	**0.74**
Cognitive Health	**0.90**	**0.91**
Emotional Health	**0.79**	**0.73**
Total	**0.94**	**0.94**
Parental Concerns	**0.86**	**0.89**

Inter-rater Reliability:	**SI Questionnaire**	**HOQ Reliability Testing**
Physical Health.....	**0.91** (0.82 – 0.96)	**0.96** (0.94 – 0.98)
Social Health.....	**0.86** (0.73 – 0.93)	**0.84** (0.73 – 0.90)
Cognitive Health.....	**0.79** (0.58 – 0.90)	**0.76** (0.61 – 0.85)
Emotional Health.....	**0.74** (0.49 – 0.87)	**0.54** (0.31 – 0.70)
Total.....	**0.89** (0.70 – 0.96)	**0.88** (0.79 – 0.93)
Parental Concerns.....	**0.78** (0.60 – 0.89)	**0.67** (0.49 – 0.79)

Overall, for each of the reliability estimates, there seemed to be a reasonable consistency in the results obtained from the SI questionnaire data and the final results obtained from the reliability testing of the final HOQ.

6.3 Discussion of Results

The sample collected for the reliability testing appeared to be appropriate in terms of its distribution of ages and etiologies. In general, it would be fair to call this sample representative of the general population of children with hydrocephalus, in whom the HOQ would be used in actual practice. Ensuring that the sample fairly reflects the population of ultimate interest is an important part of proper reliability testing.[50] The sample size used for assessing the reliability of the HOQ appeared also to be adequate and consisted of 90 completed questionnaires with 49 completed repeat questionnaires. In the majority of instances, this provided reliability coefficient estimates with fairly tight confidence intervals. Based on this sample, the HOQ demonstrated excellent TRT and internal consistency reliability with all coefficients greater than 0.80. As well, with the exception of the Social-Emotional domain, all estimates of inter-rater reliability were also greater than 0.80. Given that the goal of developing the HOQ was to enable comparison

of health status between groups of subjects (rather than individuals), these appear to be very acceptable levels of reliability[49, 151]

The Emotional domain initially demonstrated the weakest reliability. This was partly expected based on both theoretical and empirical grounds. Theoretically, this represents the least "observable" of the various health status domains. As such, the items related to Emotional health may be more difficult to answer and may be more subject to measurement error. Empirically, the results of the earlier Item Reduction testing suggested that the Emotional health domain might have a troublingly low reliability, based on the results of the SI questionnaires. In retrospect, this turned out to be exactly the case. The remedy for this was to eliminate 2 of the worst performing items and combine the Social and Emotional domains into a single domain. This had the effect of nearly doubling the number of items in the domain and greatly improving all reliability estimates. However, this combining of domains was only considered after accepting that it made theoretical sense to do so. That is, it was felt that the two domains were conceptually very similar and, in fact, the separate testing of their construct validity (see **Chapter 7**) would have proven itself to be extremely difficult. Therefore, by combining the domains, the reliability was improved to acceptable levels and the testing of construct validity was simplified, while still remaining consistent with the theoretical concepts being measured.

The reliability study involved numerous variables which could have influenced the variation in the responses to the HOQ. These included variation in location/time and respondent. The G-study demonstrated, for example, the important contribution of Child-Parent interaction to the domain of Parental Concerns. This was to be expected, of course, since this domain asks the individual parent's personal opinions about their child's health. While there would be some consistency in response based on the child's health status, there would also be expected variation based on the individual parent's psychological profile and areas of concern. Of equal importance is that Time did not appear to be a significant factor in HOQ response variation. In this study design, the variation in Time also involved a variation in location, because all Time 1 questionnaires were administered in hospital and all Time 2 questionnaires were administered at home. This gives some preliminary support to the idea that, with proper instruction, the HOQ may be able to function well as a purely mailed survey completed at the respondents' homes.

The HOQ was also tested with a small group of adolescent children (n=19). The results of this were somewhat encouraging with most domain scores having TRT and internal consistency estimates greater than 0.80. The one exception was the Social-Emotional domain with a TRT reliability of 0.60 (0.05-0.87). These results were pleasantly surprising, because the HOQ had been developed almost exclusively for parent responders. The questionnaire had not been pilot tested on children and yet, despite this, the reliability estimates appeared quite reasonable. This suggested that perhaps, amongst a properly selected sample, children of a certain age and cognitive level might be able to meaningfully complete the child version of the HOQ. However, it must be recognized

that this selected population is clearly biased towards those children with better overall health in general and better cognitive health, in particular.

When comparing the results of the child respondents to the responses of their parent, the findings were quite typical. That is, while Physical health correlated quite well, the other domains, in particular Personal Concerns, demonstrated rather poor correlations. This is consistent with much of the available literature, which suggests that higher correlations tend to be found for the more observable aspects of health (e.g., physical health) than for the less observable, more personalized aspects of health (e.g., emotional health).[152-167]

Comparison of the reliability estimates to those obtained from the severity-importance (SI) item reduction questionnaire generally suggested some consistency between the two sets of estimates. This was valuable information in at least two regards. First, it provided some retrospective justification for using the reliability estimates as guides to the selection of items during the item reduction phase. Second, it suggested that the responses to the items are quite robust. That is, for the SI questionnaire, the items were arranged fairly randomly throughout the questionnaire and not grouped into their specific domains. As well, the response options required a two step process: to assess the severity and importance for the item. In the final HOQ, the items for a given domain are located together in the questionnaire and the response option is a single step assessment of the item's severity. Yet, despite these differences, the items demonstrated quite impressive consistency of behaviour, at least as measured by inter-rater and internal consistency reliability.

6.3.1 Limitations of Reliability Testing

The method of reliability testing used in this project was subject to a number of methodological limitations. One of these was that the second questionnaire administration was performed at home via a mail survey. Although this was only done after the parents had met personally with the investigator and had completed the HOQ in hospital, it still left a number of variables unaccounted for in the completion of the second questionnaire. For example, the investigator had only rough control over the time at which the second questionnaire was completed, as was evidenced by the rather large range in the interval between completion of the two sets of questionnaires. As well, even though the investigator's pager number was given to the parents, they did not have immediate, personal access to help if they happened to have questions during completion of the questionnaire. The mailed questionnaires were clearly marked "MOTHER" and "FATHER", but it could not be absolutely guaranteed who completed the questionnaires. But perhaps the most damaging aspect of using the mail survey approach was the poor response rate. Although 66% returned the second questionnaire, only 54% were usable (that is, the questionnaires had been completed filled out without missing data). This was very disappointing, especially given the 87% response rate observed for the mailed item reduction questionnaire and the general sense of enthusiasm displayed by most parents

about this project. It is likely that the response rate was lower for this portion of the project because, on its surface, the completion of the exact same questionnaire for a second time might have seemed like a less worthwhile task to parents. This was recognized early on in the process and so during the administration of the first questionnaire, it was carefully and clearly explained to parents why completion of the second questionnaire two weeks later would be so important. The academic importance of this was stressed, but some parents mentioned that they felt as if the investigators did not trust them and wanted to "double check" their answers. Again, it was stressed as explicitly as possible by the investigator that this was not at all the case. Nonetheless, this may have at least partially explained the reduced enthusiasm for the second questionnaire.

Because the reason for lack of compliance with the second questionnaire could not be absolutely determined in most cases, it was difficult to conclude what potential biases this may have created in the final sample of those who did respond. Although examination of some basic patient characteristics, including HOQ Total Score, did not reveal substantial differences between responders and non-responders, it is reasonable to think that there might have been some systematic differences between these two groups. For example, parents who were more sure about their answers, generally less suspicious about the motive of the second questionnaire, or more interested in the project might have been over-represented amongst the responders. This, theoretically, could have resulted in artificially high reliability estimates.

Despite the limitations of the mailed questionnaire, it also provided some methodological advantage. Perhaps most importantly, by demonstrating good TRT reliability with the mailed questionnaire administration, the HOQ has proven itself to be fairly robust with the possibility that it could be useful for administration in many different types of circumstances. That is, it maybe does not need to be administered solely in hospitals. This would potentially greatly expand the uses of this instrument to include well-designed mail surveys.

6.4 Summary

The HOQ demonstrated excellent reliability properties. This included measures of internal consistency, TRT reliability, and inter-rater reliability. Almost all coefficient estimates were greater than 0.80 and would be adequate for group comparison purposes.

7. Instrument Validity

The HOQ was developed with the hope that it would be a good measure of the health status of children with hydrocephalus. The previous chapters have described the development of the questionnaire and the testing of its reliability. While the results had thus far been promising, the HOQ needed to be tested for its validity, meaning the "degree to which evidence and theory support the intended interpretations of a test or measure."[85] Throughout the earlier stages of development, attempts were explicitly made to ensure the content validity of the questionnaire. This was achieved by being comprehensive during item generation, being relatively inclusive and broad-scoped during item reduction, and by having the items reviewed by content experts. However, what remained to be tested was the HOQ's construct validity. Construct validity is the form of validity that is tested when the concept being measured is a theoretical construct that lacks a gold standard criterion against which to compare.[49] Such is the case with the measurement of health status in children with hydrocephalus. Two general forms of construct validity were tested for the HOQ: *convergent validity* (which assessed how closely the HOQ behaved compared to other, accepted measures of health) and *extreme group comparison* (which assessed how well the HOQ was able to distinguish between groups expected to differ). As well, a form of confirmatory factor analysis (CFA) was used to assess the hypothesized factor structure of the HOQ. Various types of statistical analyses were used to help in this validity testing, ranging from simple bivariate correlations to more complex structural equation modeling (SEM).

7.1 Methods of Validity Testing

The methods to test validity were essentially identical to that described for Reliability, since these were run in parallel. That is, while parents were completing the HOQ at Time 1 as described for Reliability, they also completed the instruments that would be used for testing of construct validity (see below). As well, part of the follow-up mailing for Time 2 involved repeating some of the instruments (those that did not require direct administration by the investigator). The sample size requirement for the validity testing was based primarily on the needs of the SEM. This was estimated to be a minimum sample size of 50 subjects (see **Section 7.1.2**).

7.1.1 Instruments Used to Test Construct Validity

Construct validity relies on testing the relationship between the new instrument and other parameters. It is hypothesized that, if the new instrument is truly measuring a certain concept, it will have a predictable relationship with another, proven measure of a

related concept. The following measures were used in the validity testing of the new instrument. For each instrument, a description is followed by the hypothesized Pearson correlations (low is <0.3, mod is 0.3-0.6, high is >0.6).

7.1.1.1 Functional Independence Measure for Children (WeeFIM)

This measure was developed in 1994 to serve as a "measure of disability" to be used across different settings in the pediatric population.[143] It measures the degree of performance of certain tasks. That is, it is a measure of disability rather than impairment, as defined by the World Health Organization.[117] It consists of 18 items, each representing certain tasks, covering 6 domains: Self-care, Sphincter control, Physical transfers, Locomotion, Communication, and Social Cognition.(see **Appendix G – Functional Independence Measure for Children**) For each of the items, the parent is asked to rate the child's degree of independence on a 7-point scale, with 7 being complete independence and 1 being requiring total assistance. This scale has been used in children with varying types of illnesses, including limb deficiency, Down's syndrome, spina bifida, and cerebral palsy.[143] These children have ranged in age from 6 months to 14 years. Throughout these uses, the WeeFIM has shown excellent measurement properties (test-retest correlation 0.89-0.99, inter-rater correlation 0.80-0.96). The domains within the WeeFIM can be combined to form subscales, as was described by Msall et al..[143] These subscales include: Self-care/sphincter control, Transfers/locomotion, Communication/social cognition.

Based on the data collected from our sample, there appeared to be high correlation among all the subscale scores (Pearson >0.63) and the WeeFIM Total Score (Pearson >0.85). Therefore, only the Total Score was used in hypothesizing correlations with the HOQ:

TABLE 7.1. WeeFIM hypotheses

WeeFIM	HOQ Domains			
	Physical	Social-Emotional	Cognitive	Total
Total WeeFIM	*high*	mod	mod	*high*

The purpose of using the WeeFIM was to test the construct validity of three of the new instrument domains (Physical, Social-Emotional, Cognitive) and the total instrument score. The completion of the WeeFIM required the help of the investigator, so this was only administered at Time 1.

7.1.1.2 Strengths and Difficulties Questionnaire (SDQ)

The SDQ was developed by Goodman in 1998.[145] The purpose of this was to

serve as a screening tool for the "behaviour, emotions, and relationships" of children. It was essentially an abbreviated version of the Rutter parent questionnaire and was meant to be used in children aged 4 to 16 years. The SDQ consists of 25 items, representing both positive and negative attributes.(see **Appendix H – Strengths and Difficulties Questionnaire**) The parents respond to the attributes on a 3-point scale as either "not true", "somewhat true", or "certainly true". The results of the SDQ can be used to calculate five scores: Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems, and Prosocial Behaviour. These dimensions were determined by the developer using factor analysis of the expanded Rutter parent questionnaire.[145] The SDQ has demonstrated high correlations with the full Rutter questionnaire (ranging from 0.78-0.88 for the subscores).[145] The test-retest reliabilities have been reported as 0.85 for the Total Score, and between 0.70-0.85 for the subscores.[168]

Based on our data sample, the correlation between the some of the SDQ subscales (Emotional, Conduct, and Hyperactivity) and the Total Score were high (Pearson >0.65). Therefore, only the Total Score was used in place of these subscores. It was hypothesized that the follow correlations between the SDQ and the HOQ would be observed:

TABLE 7.2. SDQ hypotheses

SDQ Subscales	HOQ Domains			
	Physical	Social-Emotional	Cognitive	Total
Peer Problems	low	mod	low	mod
Prosocial Behaviour	low	mod	low	mod
Total SDQ	mod	***high***	mod	mod

The SDQ primarily allowed for the testing of the construct validity of the Social-Emotional domain of the new instrument.

7.1.1.3 Health Utilities Index-Mark 2 (HUI-2)

The HUI-2 is a multiattribute health status classification system and is based on the concept that health consists of a number of different attributes and within each of these there are several different levels of function.(see **Appendix I – Health Utilities Index - 2**) These levels aim to reflect stepwise increments between good and poor functioning.[169] This system has taken various modified forms and has been used in different pediatric populations, demonstrating good reliability and validity.[92, 170, 171] The original form was described by Torrance et al. in 1982.[90] The HUI-2 assesses a child's health status in the following attributes: sensation, mobility, emotion, cognition, self-care, and pain. An advantage of the HUI-2 is the ability to translate any individual health state

into a utility score that estimates the relative preference for that state.[172]

Based on our sample data, there was high correlation between some of the subscale scores (Sensation, Mobility, Cognition, Self-Care) and the Total HUI Score (Pearson >0.68). These subscales were, therefore, replaced with just the Total Score in predicting correlations. The following were the hypothesized correlations between the HUI-2 and the HOQ:

TABLE 7.3. HUI-2 hypotheses

HUI-2 Subscales	HOQ Domains			
	Physical	**Social-Emotional**	**Cognitive**	**Total**
Emotion	low	***high***	low	mod
Pain	low	low	low	low
Total HUI-2	mod	mod	mod	***high***

Since the HUI-2 required the help of the investigator for completion, this was administered only at Time 1.

7.1.1.4 Impact-on-Family Scale (IFS)

The IFS was developed by Stein and Riessman in 1980.[144] The measure attempts to quantify the impact of childhood illness on a family, i.e., any change in the normative behaviour of the family which is attributable to the child's illness. This is a 24-item questionnaire with a 4-point Likert response scale.(see **Appendix J – Impact-on-Family Scale**) Factor analysis was used to categorize the items into 4 dimensions: Financial Burden, Familial/Social Impact, Personal Strain, and Mastery. Subscores can be calculated for each dimension, along with a total overall score. The internal consistencies (alpha) for the individual dimensions range from 0.60 to 0.86 and 0.88 for the total score.[144] Others have shown moderate correlation of the IFS to maternal mental health in mothers of children with chronic illnesses,[173] to parental quality of life for the parents of children with asthma,[174] and to indicators of health status and quality of life in children with myelomeningocele[135] and cystic fibrosis.[175]

Based on our sample data, there was high correlation (Pearson >0.85) among all the subscores of IFS and the total score, with the exception of Mastery, which had near zero correlation with the others. This was expected, since it represents a positive attribute, rather than a negative attribute like the others. Therefore, only the total IFS score and the Mastery subscore were used in predicting correlations. The following were the hypothesized correlations between the IFS and the HOQ:

TABLE 7.4. IFS hypotheses

IFS Scales	HOQ Domains				
	Physical	Social-Emotional	Cognitive	Total	Parental Concerns
Mastery	low	low	low	low	low
Total IFS	***high***	mod	mod	***high***	***high***

The IFS was selected primarily to provide construct validity for the Parental Concerns domain.

7.1.1.5 Wide Range Achievement Test – 3 (WRAT-3)

The WRAT was developed originally in the 1930's by Joseph Jastak.[176] Since that time, it has undergone several revisions and been used to assess several thousand subjects. The latest revision resulted in WRAT-3 which consists of 3 subtests: Reading, Spelling, and Arithmetic. Normative data is available for each of these subtests and raw scores for children 5 years and older can be converted to standard scores to allow for comparisons across age ranges. For this project, only the Reading subtest was used, since it required the least physical ability from the children (the other subtests mandate a written component). Scores on the Reading subtest are highly correlated with Spelling subtest scores (Pearson correlation 0.87), rendering some redundancy between these two subtests.[176] As well, the Reading subtest has demonstrated the highest test-retest reliability (0.93) of all the subtests and excellent internal consistency (alpha 0.91).[176] It was hypothesized that the WRAT Reading test would demonstrate high correlation with the HOQ Cognitive domain. Since the WRAT-3 was administered by the investigator, this was performed only at Time 1.

7.1.2 Structural Equation Modeling (SEM)

Structural equation modeling was used in two separate ways to test hypotheses about the HOQ: it was used to test the hypotheses regarding construct validity and it was used to test the hypothesis of the proposed multidimensional factor structure of the HOQ, i.e., confirmatory factor analysis. These hypotheses were tested with separate models. For each of the models, the fit was assessed using various statistical tests:

Absolute fit: χ^2 test
Relative χ^2 (chi-squared/ degrees of freedom)[177]

Relative fit: Tucker-Lewis index (TLI)[178]
Bollen relative fit index (RFI)[179]
Comparative fit index (CFI)[180]
Parsimony-adjusted normed fit index (PNFI)[177]

The claimed advantage of the first three relative fit indices is their relative independence to sample size. This has been empirically shown to be the case for the TLI and CFI, but not for the Bollen RFI.[181, 182] The PNFI is used to adjust for model complexity, penalizing more complex models. All SEM analyses were performed using AMOS 4.01 (Smallwaters Corp., Chicago, Illinois, USA) with maximum likelihood estimation.

Sample size requirements for SEM have not been well established, although it is generally accepted that larger sample sizes are better.[183] Zwick and Velicer recommended at least 2 to 5 subjects per variable in the model.[184] Since the most complex model being used involved 24 unknown parameters (see below), the required sample size was felt to be between 48 and 120 subjects.

Note: For all the following SEM diagrams, the mean value of unobserved variables (including "error") was assumed to be "0" (as is labeled in the diagrams), any regression weight which was fixed is indicated by the value "1", all other parameters that were constrained to be equal were labeled with the same letter, e.g., "a" or "b". All free parameters were left unlabeled.

7.1.2.1 SEM to Confirm Multidimensional Factor Structure

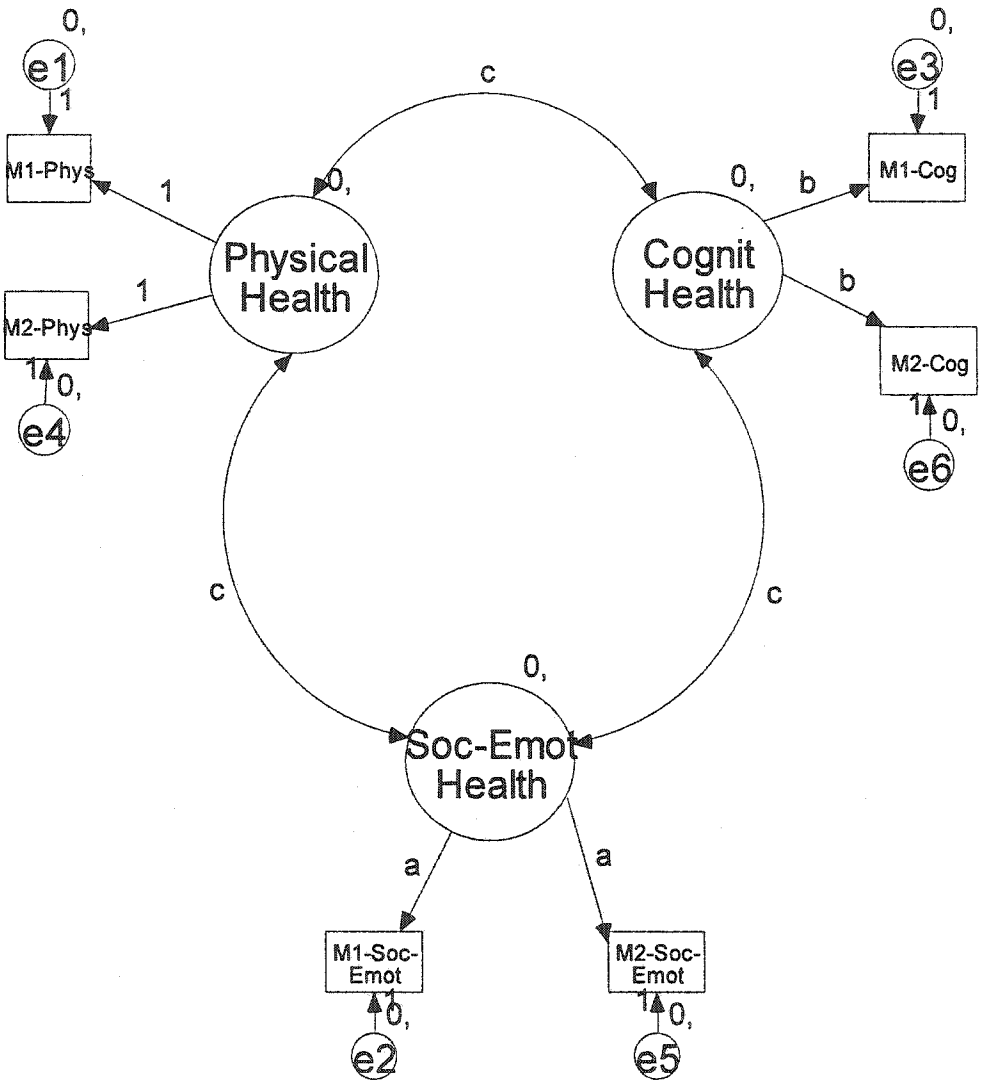
In order to confirm the multidimensional nature of the HOQ, the fit of two competing models was compared.[185] The first model (see **Model A**) assumed multidimensionality, with the three latent factors being Physical Health, Social-Emotional Health, and Cognitive Health. Each latent factor was represented by two indicator variables. The indicator variables were determined by splitting the HOQ items within each domain into two groups by taking alternate questions (i.e., one group contained all even numbered items and the other contained all odd numbered items). The response scores for the items were summated to create two indicator variables for each domain. For example, "M1-Phys" represented the summated score from all odd numbered items in the Physical Health domain and "M2-Phys" represented the summated score for all even numbered items in the Physical Health domain. The second model (see **Model B**) assumed unidimensionality, with only one latent factor: Overall Health. The same six indicator variables were used.

7.1.2.2 SEM to Confirm Construct Validity

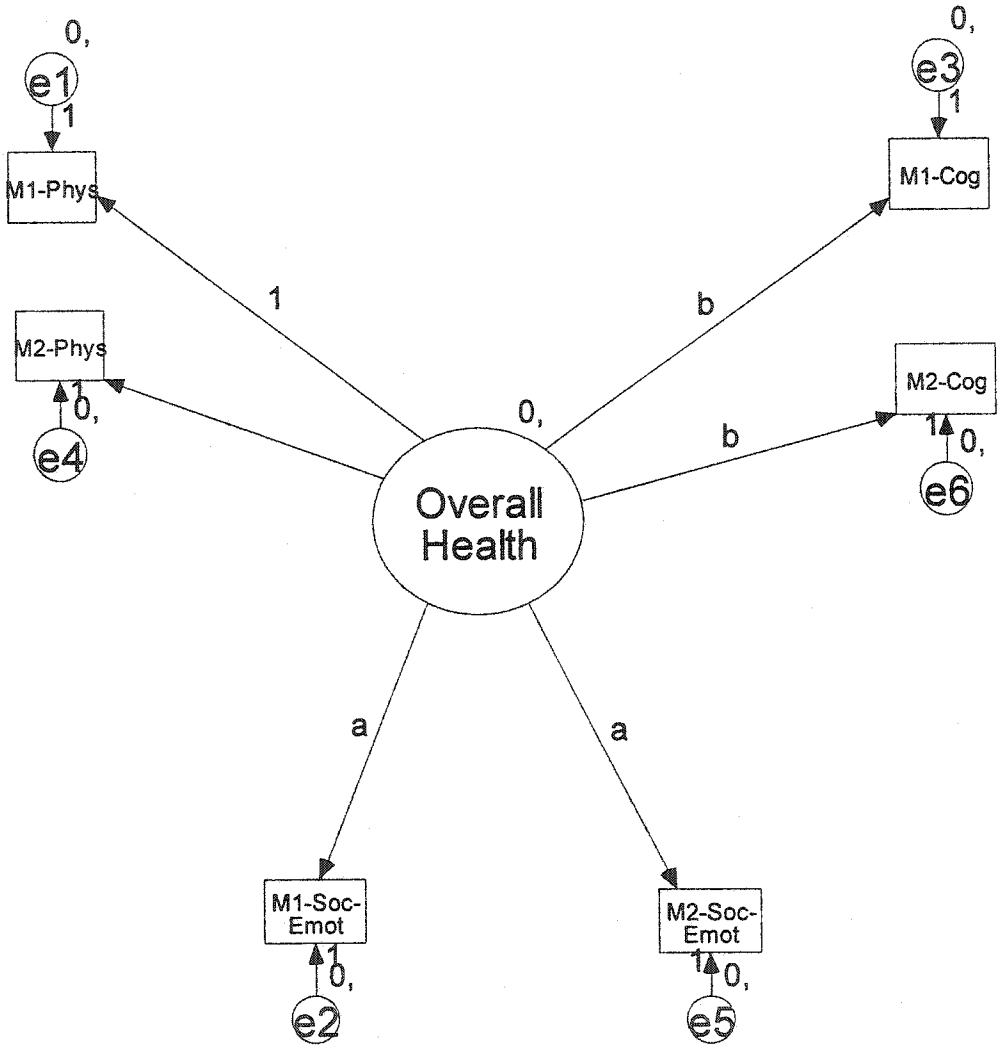
Model C displays graphically the SEM used to test the hypotheses related to construct validity. This represented a simplified form of the various hypotheses presented in the previous section and highlights the expected relationship amongst the observed

HOQ variables and the observed variables used to test construct validity, i.e., WeeFIM, WRAT, SDQ.

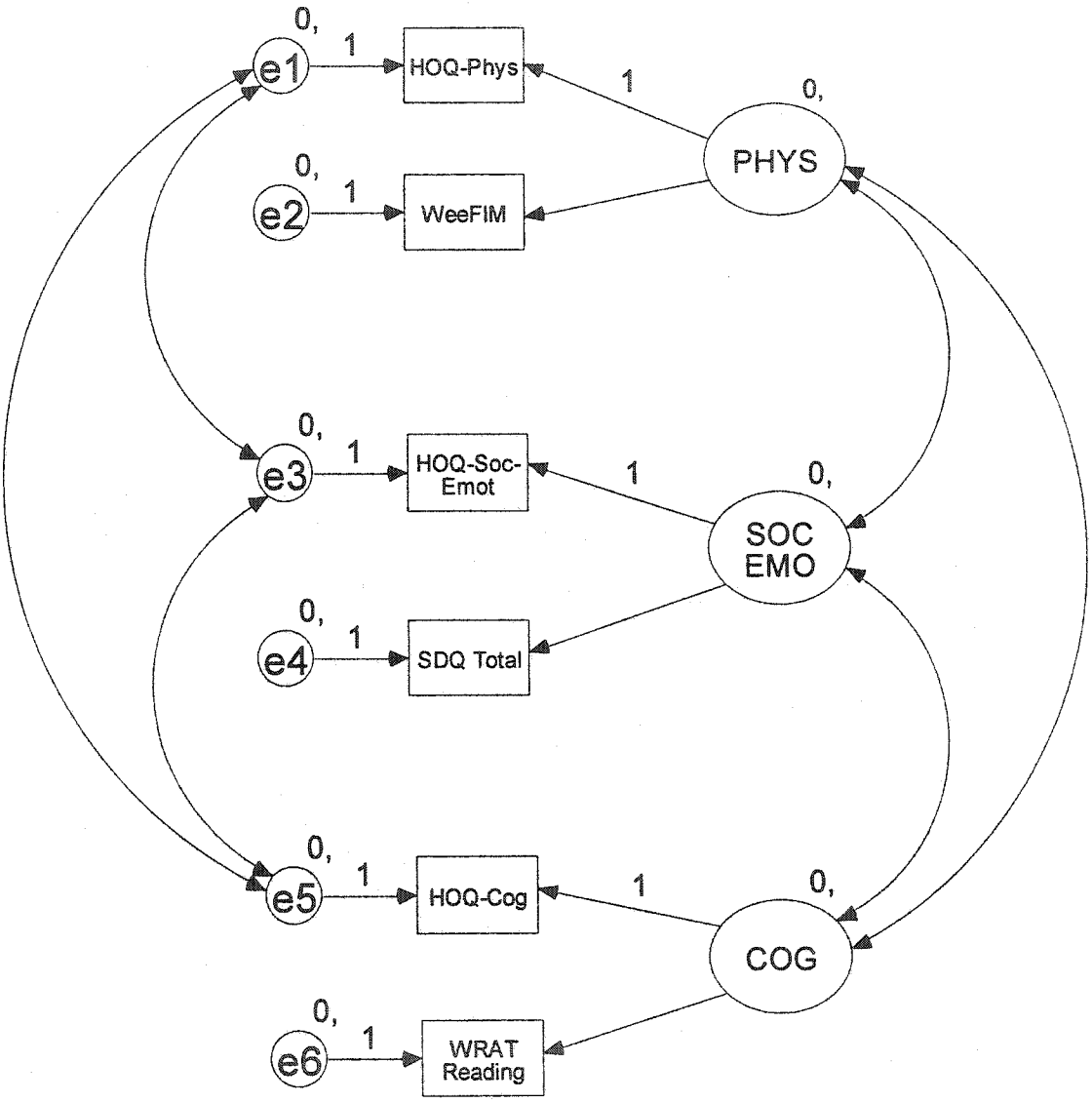
FIGURES 7.1 to 7.3. (on the following three pages) Structural equation models A, B, and C



Model A. Multidimensional Factor Model



Model B. Unidimensional Factor Model



Model C. Construct Validity Model

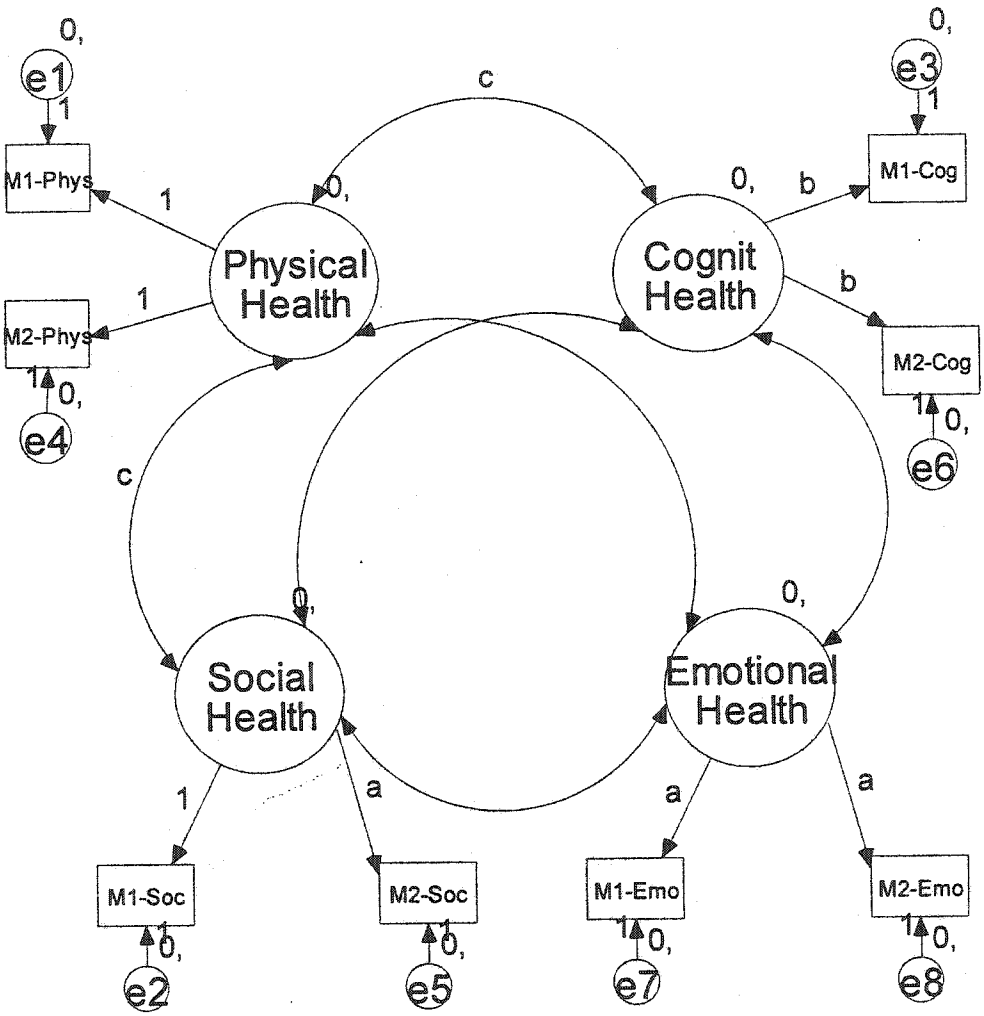
7.1.3 Verifying the Acceptability of the Social-Emotional Domain

During the Reliability testing, the decision was made to collapse the Social and Emotional domains into a single domain (Social-Emotional health). The rationale for this was discussed in **Section 6.2.3,** and was based predominantly on the improvement in the reliability estimates of the combined domain. However, part of the justification for this was that the two domains were very similar conceptually. In order to test this assumption, two types of analyses were carried out:

i. The hypothesized correlations described in the Construct Validity section (**Section 7.1.1**) were rechecked with the original Social and Emotional domains. It was hypothesized that the correlations of these two domains would be similar for the majority of the tests (i.e., the difference in the Pearson correlations would be within ±0.3 – which is the range of a level of magnitude between low, moderate, and high correlations as described in **Section 7.1.1**).

ii. Another SEM (**Model D**) was constructed which tested the multidimensional structure of the HOQ, but this time included separate domains for Social and Emotional health. The fit indices for this model were tested against those for the other multidimensional model (**Model A**), as described in **Section 7.1.2**. It was hypothesized that there should be no significant difference between the two models.

FIGURE 7.4. (on following page) Structural equation model D



Model D. Multidimensional Model - Alternate

7.1.4 Extreme Group Comparisons

Most of the above methods of testing validity deal with *convergent* construct validity, as defined by Tugwell and Bombardier.[47] Convergent construct validity involves assessing how well a new instrument agrees with well-accepted, related measures. *Extreme group comparison* involves assessing whether an instrument is able to show differences between groups that one would expect to differ in the concept being measured. Based on the available hydrocephalus literature and on the opinion of experts, the following hypotheses were tested:

> i. Children with shunt infections would have a worse measured health status than those without shunt infection, particularly in the domain of Cognitive health.[186]
>
> ii. Children with one or fewer shunt revisions would have a better measured health status than those with two or more shunt revisions, particularly in the domain of Cognitive health.[187]
>
> iii. Children with epilepsy would have a worse health status than those without epilepsy, particularly in the domains of Cognitive and Social-Emotional health.[188]

In addition to the above univariable analyses, a multivariable ANOVA was performed using the HOQ Total Score as the dependent variable and the following as independent variables:

Shunt infection (yes or no) – fixed factor
Shunt revisions (≤ 1 or ≥ 2) – fixed factor
Epilepsy (yes or no) – fixed factor
Etiology (4 separate categories used) – fixed factor
Age (in years) – continuous variable

The significance of each variable was assessed using the F-test statistic. Two separate models were tested: one without any interaction terms and another full factorial model with all possible interaction terms included.

7.2 Results of Validity Testing

7.2.1 Construct Validity

The following tables represent the Pearson correlations obtained between the

HOQ and the other independent health measures. Highlighted, italicized values indicate those which had been hypothesized, *a priori*, to have a high correlation (>0.60). When such values were greater than 0.60, a p-value was calculated using a single-sided t-test to assess whether the value was significantly greater than 0.60:

TABLES 7.5 to 7.8. Pearson correlations with the HOQ

WeeFIM Subscales	HOQ Domains			
	Physical	Social-Emotional	Cognitive	Total
Total WeeFIM	*0.89 (p=0.004)*	0.43	0.47	*0.73 (p=0.13)*

SDQ Subscales	HOQ Domains			
	Physical	Social-Emotional	Cognitive	Total
Peer Problems	0.18	0.49	0.27	0.37
Prosocial Behaviour	0.28	0.31	0.20	0.32
Total SDQ	0.36	*0.74 (p=0.11)*	0.59	0.66

HUI-2 Subscales	HOQ Domains			
	Physical	Social-Emotional	Cognitive	Total
Emotion	0.13	*0.45*	0.27	0.34
Pain	0.22	0.13	0.15	0.19
Total HUI-2	0.88	0.56	0.57	*0.81 (p=0.03)*

IFS Scales	HOQ Domains				
	Physical	Social-Emotional	Cognitive	Total	Parental Concerns
Mastery	0.05	0.00	0.03	0.03	0.07
Total IFS	*0.71 (p=0.17)*	0.36	0.36	*0.58*	*0.59*

In addition, the expected high correlation between the WRAT Reading test and the HOQ Cognitive domain was, in fact, 0.59.

7.2.2 SEM to Confirm Multidimensional Factor Structure

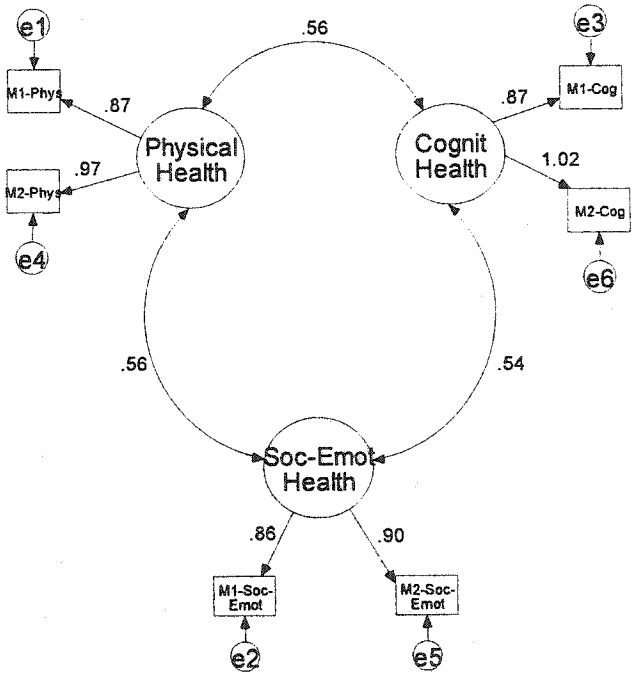
The sample size for these analyses consisted of the 80 patients whose mothers completed the HOQ at Time 1. All other cases were excluded from the analysis. The results of the SEM analysis of **Models A** and **B** are presented below:

TABLE 7.9. Fit indices for SEM

	Model A: Multidimensional Model	Model B: Unidimensional Model
Chi-squared test of absolute fit ($\chi 2$, $\chi 2$ / df, p-value)	24.39, 2.71, p=0.004	152.92, 13.90, p<0.001
Tucker-Lewis index of relative fit	0.97	0.80
Bollen relative fit index	0.96	0.79
Comparative fit index	0.99	0.90
Parsimony-adjusted Normed fit index	0.42	0.47

The multidimensional model appeared to have a much better statistical fit than the unidimensional model. By applying a chi-squared difference test between the two models, the resulting statistic is $\chi^2 = 128.53$, df=2, and $p<0.001$, suggested a significant difference in favour of the multidimensional model.[189] However, in absolute terms the fit of the multidimensional model was good, but not excellent. The estimates of the significant standardized regression weights (between latent and observed variables) and correlations (between latent variables) of **Model A** are shown below:

FIGURE 7.5. Estimates for Model A



Model A. Multidimensional Factor Model wit Parameter Estimates.

7.2.3 SEM to Confirm Construct Validity

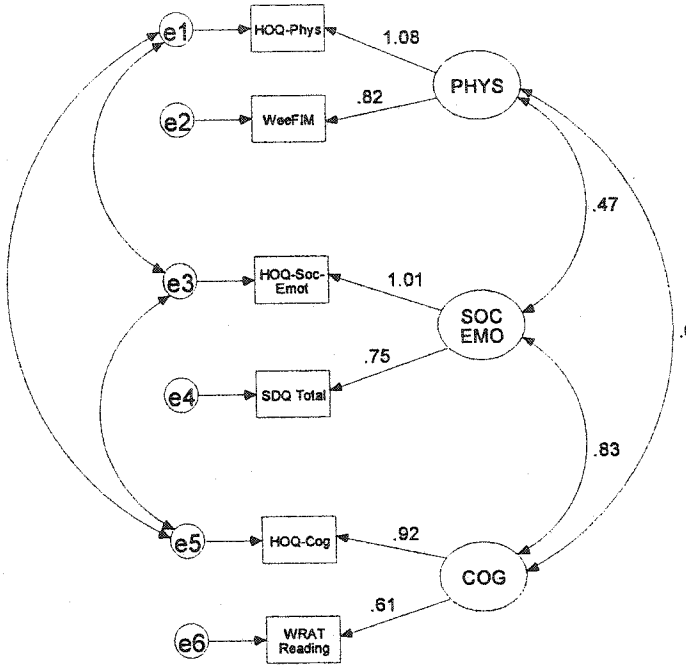
The sample used for this analysis consisted of 80 patients whose mothers completed the Time 1 administration of the HOQ. All other cases were excluded from the analysis. The results of the SEM analysis of **Model C** are presented below:

TABLE 7.10. Fit indices for SEM

	Model C: Construct Validity
Chi-squared test of absolute fit ($\chi 2$, $\chi 2$ / df, p-value)	8.13, 2.71, p=0.04
Tucker-Lewis index of relative fit	0.98
Bollen relative fit index	0.96
Comparative fit index	0.99
Parsimony-adjusted Normed fit index	0.14

The results indicated a very good relative fit of the proposed model, adding further evidence to the construct validity of the HOQ. The absolute fit test, however, was marginally significant. The estimates of the significant standardized regression weights (between latent and observed variables) and correlations (between latent variables) of **Model C** are shown below:

FIGURE 7.6. Estimates for Model C



Model C. Construct Validity Model with Correlation and Covariance Estimates

7.2.4 Acceptability of the Social-Emotional Domain

The correlations between the tests of construct validity and the original Social and Emotional domains, along with the Social-Emotional combined domain are shown below:

TABLE 7.11. Pearson correlations with the Social and Emotional domains

	Social Domain	Emotional Domain	Social-Emotional Domain
Wee FIM Total	0.53	0.19	0.43
SDQ – Peer Problems	0.59	0.23	0.49
SDQ – Prosocial Behaviour	0.27	0.25	0.31
SDQ Total	0.58	0.68	0.74
HUI - Emotional	0.24	0.53	0.45
HUI - Pain	0.15	0.08	0.13
HUI Total	0.60	0.34	0.56
IFS - Mastery	0.00	0.00	0.00
IFS Total	0.48	0.12	0.36
WRAT Reading Test	0.47	0.26	0.44

Three of the 10 comparisons had differences between the correlations of greater than 0.3, while the majority had differences less than 0.3. As well, the mean difference in the correlations was 0.12 (95% confidence interval 0.00 – 0.28, $p = 0.11$ using paired-sample t-test).

The results of the SEM analysis comparing **Models A** and **D** showed very similar fit indices:

TABLE 7.12. Fit indices for SEM

	Model A: Multidimensional Model	**Model D: Multidimensional Model - Alternate**
Chi-squared test of absolute fit ($\chi 2$, $\chi 2$ / df, p-value)	24.39, 2.71, p=0.004	46.12, 2.71, p<0.001
Tucker-Lewis index of relative fit	0.97	0.96
Bollen relative fit index	0.96	0.94
Comparative fit index	0.99	0.98
Parsimony-adjusted Normed fit index	0.42	0.46

7.2.5 Extreme Group Comparisons

The HOQ Scores for children with and without previous shunt infection were compared using independent samples t-test. The table below shows the mean values ± standard error of the mean:

TABLE 7.13. HOQ scores relative to shunt infection

HOQ Domain	No Shunt Infection (N=64)	Shunt Infection (N=16)	P-value (t-test)
Physical	0.74 ± 0.03	0.59 ± 0.07	***0.05***
Social-Emotional	0.72 ± 0.02	0.73 ± 0.04	0.82
Cognitive	0.60 ± 0.03	0.43 ± 0.07	***0.03***
Total	0.70 ± 0.02	0.62 ± 0.04	0.07
Parental Concern	0.47 ± 0.03	0.36 ± 0.08	0.19

There were significant differences within the domains of Physical and Cognitive Health, but only a trend toward a difference in the Total Score.

The HOQ Scores for children with one or fewer shunt revisions were compared to those with two or more revisions using independent samples t-test. The table below shows the mean values ± standard error of the mean:

TABLE 7.14. HOQ scores relative to shunt revisions

HOQ Domain	≤ 1 Shunt Revision (N=53)	≥ 2 Shunt Revisions (N=27)	P-value (t-test)
Physical	0.74 ± 0.03	0.65 ± 0.05	0.14
Social-Emotional	0.74 ± 0.02	0.69 ± 0.03	0.20
Cognitive	0.63 ± 0.03	0.44 ± 0.05	***<0.01***
Total	0.72 ± 0.02	0.62 ± 0.03	***0.02***
Parental Concern	0.49 ± 0.04	0.36 ± 0.05	***0.04***

Significant differences were demonstrated for Cognitive Health and Total Score, as well as for Parental Concern.

The HOQ Scores for children without epilepsy (defined as a response of "Not at all true" to item 13: "My child has many seizures") were compared to those with epilepsy (defined as all other patients) using independent samples t-test. The table below shows the mean values ± standard error of the mean:

TABLE 7.15. HOQ scores relative to epilepsy

HOQ Domain	No Epilepsy (N=64)	Epilepsy (N=16)	P-value (t-test)
Physical	0.78 ± 0.02	0.43 ± 0.07	***<0.0001***
Social-Emotional	0.75 ± 0.02	0.62 ± 0.03	***0.001***
Cognitive	0.61 ± 0.03	0.39 ± 0.06	***0.002***
Total	0.73 ± 0.02	0.51 ± 0.03	***<0.0001***
Parental Concern	0.49 ± 0.03	0.27 ± 0.06	***0.003***

All scores showed statistically significant differences between the two groups.

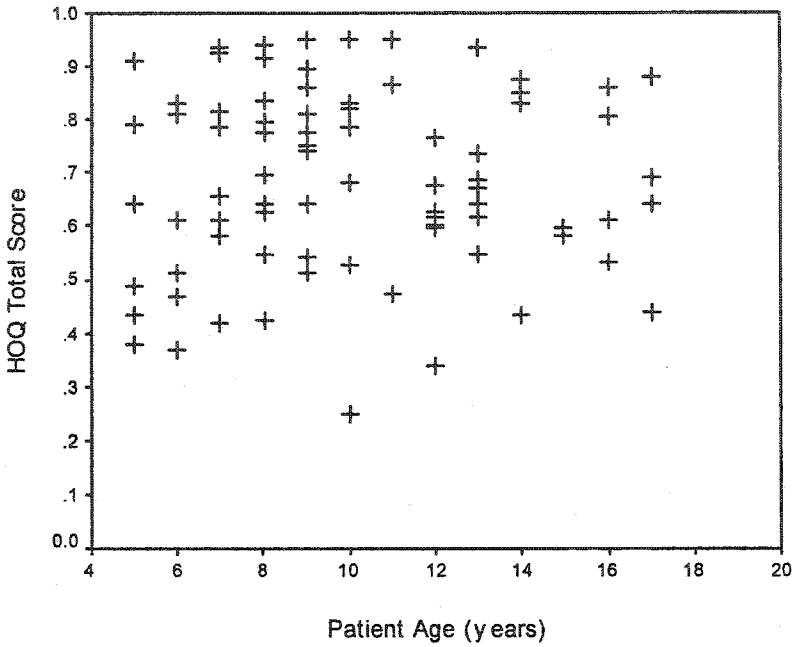
In order to account for possible confounding in the above relationships, an exploration of the relationship between several independent variables and the dependent variable of HOQ Total Score was carried out using ANOVA. The following independent variables were included in the multivariable model, with their respective p-values (from F-test):

	Beta Coefficient	Standard error	P-value
Shunt infection:			
no	0.01	0.04	0.73
yes	0.00		
Shunt revisions:			
≤ 1	0.06	0.04	0.13
≥ 2	0.00		
Epilepsy:			
yes	0.23	0.04	<0.0001
no	0.00		
Etiology:			
congenital	0.04	0.06	0.50
spina bifida	-0.08	0.06	0.22
others	-0.02	0.05	0.72
hemorrhage	0.00		
Age (in years)	0.006	0.005	0.20

The R^2 for the model was only 0.35, suggesting that there was still a large amount of variation unaccounted for. When a full factorial model was tested, in which all possible interaction terms were included in the model, none of the other individual variables were significant, while the p-value for Epilepsy remained <0.0001. The R^2 for the full factorial model improved to 0.49.

Of note in this analysis was that age did not seem to be related to the HOQ Total Score. The possibility of an age-related effect had been a concern during the early development of this questionnaire. To confirm this lack of relationship, the following scatter-plot was analyzed:

FIGURE 7.7. Scatter-plot of HOQ scores relative to age



A linear regression analysis of this data showed that the β coefficient for Age was essentially zero (β = 0.003, 95% confidence interval = -0.009 to 0.014, p-value 0.65). Similar analyses confirmed the lack of relationship between Age and each of the domain scores (Physical, Social-Emotional, Cognitive, and Parental Concern).

7.3 Discussion

7.3.1 Construct Validity

It is somewhat difficult to summarize the results of the construct validity testing, since it involved a myriad of hypothesized correlations. In general, the following could be said:

i. The Physical domain demonstrated excellent correlations with the WeeFIM scores, as expected. The correlation of the WeeFIM scores with the other domains was only moderate at best.

ii. Correlations of the Social-Emotional domain with the SDQ scores were mostly moderate. This may partially be explained by the psychometric properties of the Social-Emotional domain scores, which showed lower (although still acceptable) levels of reliability (see **Chapter 6**). As well, the social and emotional aspects of health are probably the most vaguely defined and it is perhaps expected that two different measures

of these will be more different than two separate measures of physical health, for example. Regardless, although the correlations were not as high as expected, they were generally higher than the correlations between the SDQ and the other HOQ domains. This was at least somewhat comforting.

iii. The Cognitive domain and the WRAT Reading test were very nearly highly correlated (0.59), as had been expected.

iv. The overall HUI-2 utility score demonstrated a very high correlation with the HOQ Total Score (0.81). This would prove to be very valuable in establishing the interpretability of the HOQ (see **Chapter 0**).

v. Although the IFS correlated reasonably well with the HOQ Parental Concerns domain, the correlations were never very high and were somewhat worse than the correlation with the HOQ Physical domain. The concept of parental concern is a much harder one to test than the others, primarily because it is not one that has been extensively investigated before. Therefore, finding established measures against which to compare was difficult. The IFS was the closest one that could be found, but even it did not address the exact same concept that was being tapped by the HOQ Parental Concerns domain.

7.3.2 Extreme Group Comparisons

The differences between the various groups were largely as had been hypothesized:

i. Children without shunt infection had better Physical and Cognitive scores.
ii. Children with fewer than 2 shunt revisions had better Physical, Cognitive, Total and Parental Concern scores.
iii. Children without epilepsy had better scores in all domains.

However, in a multivariable analysis, the overwhelmingly significant variable was the presence of epilepsy, accounting for the most variation in the HOQ Total Score. The other variables no longer showed any statistical significance. This finding is consistent with other studies which have looked at the impact of epilepsy on children with hydrocephalus.[21, 188] However, a recognized limitation of the use of this "epilepsy" variable is that it is derived from a component of the HOQ itself. That is, the presence or absence of epilepsy was based on the parent's response to Item #13: "My child has many seizures". There were at least two potential methodological problems with this approach. First, the assumption that an answer of "Not at all" corresponded to "no epilepsy" may not have been correct, depending on how the parent interpreted the item and its response. Second, since the item was part of the questionnaire, it undoubtedly had some inherent correlation to the HOQ Total Score. Minimizing this effect, however, was the fact that it represented only one of 51-items, so its contribution to the Total Score was actually quite

small.

It is interesting that, although epilepsy proved to be such a substantial variable, the item related to epilepsy was almost not included in the HOQ. In fact, it was initially eliminated and then re-instated based on the recommendation of the content experts.(see **Section 5.2.1.4**) This points to, perhaps, a weakness in the item reduction methodology. Why would an item which appears to be so strongly associated with the concept being measured be nearly eliminated? The methodology used in item reduction required that a relatively high proportion of the sample be affected by the item and that it be rated as very important to those subjects. Epilepsy was present in only a minority of the Validity sample, so it is possible that this might have resulted in a relatively lower severity-importance score during item reduction. On the other hand, even if the epilepsy item were not included in the final HOQ, the total scores would likely still be rather similar and the relationship to the presence of epilepsy would likely still have been demonstrated. In other words, the validity of the overall HOQ score would remain intact. Alternatively, one could interpret this as highlighting the importance of using content experts in the development of health status measures.

7.3.3 Structural Equation Modeling

7.3.3.1 Factor Structure

In order to confirm the hypothesized multidimensional factor structure of the HOQ, there were numerous options available. Broadly, one could have performed an exploratory factor analysis (EFA) to identify the statistical factors that presented themselves and then attempt to reconcile these with the hypothesized structure. One of the major criticisms of EFA, however, is its relative lack of objectivity; that different researchers could come up with very different models from the same data set and that they could all be statistically sound and justified.[190] Much of this potential for lack of objectivity comes from the fact that many of the crucial steps involved in EFA do not come with absolute rules of conduct. For example, deciding how many factors to retain, or what level constitutes a strong factor loading, and the use of rotation adjustment are relatively arbitrary decisions.[190] As well, even after one has developed a statistical model of factors and their respective items, it remains a large jump from this statistical creation to one of factor labeling and explanation.

Confirmatory factor analysis (CFA), unlike EFA, begins with a hypothesis about the theoretical factor structure and proceeds to testing of this hypothesis. That this is more akin to deductive logic makes CFA, perhaps, somewhat more appealing to clinical researchers. Nonetheless, CFA is also subject to criticism. Probably the most important limitation of CFA is the lack of unanimity in the statistical testing of the hypothesis, which, of course, is the key feature of CFA. There is no consensus on even which broad type of test to use: tests of absolute fit or tests of relative fit. There are theoretical problems with using tests of absolute fit, namely that in most cases, the assumption of

multivariate normality is not met.[98] As well, one begins with a null hypothesis which states that the model is a perfect fit of the data. However, it is generally accepted that the model is only an approximation, so we know the null hypothesis to be false *a priori*. Tests of relative fit do not aim to test for perfect fit, but, rather, reflect how much better the proposed model fits the data compared to a null model in which there are no underlying factors. The problem with these tests lies in their interpretation – it is not clear at what level a relative fit should be considered acceptable.[98]

Accepting all the previously discussed limitations, it was decided that for the HOQ, the proposed factor structure would be tested by CFA, using the techniques of SEM. This still left many options, especially with deciding on what the final model would look like and how would one statistically test for its acceptability. The model that was decided upon (Model A) is a simplified model that attempts to reflect the hypothesized factor structure. The three factors were allowed to correlate, as would be expected given that they are all presumed to represent a facet of health. Each of the three factors was given 2 indicator variables by grouping together the odd and even numbered items within each domain. The alternative to this would have involved using each questionnaire item as a separate indicator. However, this would have led to an enormously large and complex model (with over 50 indicator variables) and it was not felt that the available sample size would have been able to sufficiently test such a model. Because of the lack of consensus regarding the statistical testing of models, no one test was relied upon to show acceptability. Instead, several measures were used and they were used in a relative comparison to Model B, which hypothesized a single factor (Overall Health) only. Using this method of testing, the multidimensional model (Model A) demonstrated very good indices of relative fit (all >0.90), although the chi-squared test of absolute fit was significant. However, the fit of the multidimensional model was substantially better than the unidimensional model (Model B) which demonstrated poorer indices of relative fit (most <0.85) and a much higher chi-squared statistic (also significant). As well, the relative chi-squared statistic for Model A was only 2.71 compared to 13.90 for Model B. While the interpretation of this statistic is still open to debate, values under 5.0 have been considered as indicators of acceptable model fit.[177, 191] The PNFI was similar for the two models, presumably because it penalized the multidimensional model for added complexity. Regardless, the overall interpretation of the results was that, compared to a unidimensional model, the multidimensional model was a much better fit for the data collected, thus confirming the hypothesized factor structure.

One of the major limitations of this SEM is the relatively small sample size (n=80). Although the criteria for adequate sample size for SEM are not well-established, one rough guideline is to aim for approximately 4 subjects for every unknown parameter in the model .[189, 192] For these two models, the number of unknown parameters was 16 and 18, rendering the actual sample size as possibly acceptable. The potential problems of a small sample size were recognized *a priori* and influenced the choice of statistical tests used in the analysis of the model. Specifically, measures of relative fit that are

believed to be relatively independent of sample size were chosen. However, it is likely that all the tests used were influenced, to some degree, by the small sample size. The chi-squared statistic has been shown empirically to be falsely elevated in the presence of small sample sizes, e.g., n=50.[193] This would lead to an erroneous rejection of the model, i.e., an elevated Type I alpha error. This may partially explain the rejection of both SEM based on the significant chi-squared tests. The relative fit indices which were used probably are also affected by sample size to varying degrees. The direction of the effect appears to be similar as that for the chi-squared test, in so much as smaller sample sizes tend to result in higher rates of falsely rejecting the model.[193] Regardless of this potential effect, however, all the relative fit indices suggested a good fit for the multi-dimensional model.

As further protection against the potential bias of a small sample size, none of the fit indices were used in isolation. That is, the analysis involved assessing the relative fit of two clearly defined, competing models with each having the same sample size. This effectively equilibrated the small size biases and allowed for a more meaningful comparison.

7.3.3.2 Construct Validity

The results of the construct validity testing presented earlier suggested that the HOQ largely demonstrated the expected relationship with the other measures of health. However, to confirm this in a more unified format, SEM was used. The model (**Model C**) was constructed using the multidimensional factor structure suggested by the CFA. Each of the three factors was represented by two indicator variables: the domain score from the HOQ and the separate health measure used for construct validity testing. While this ignored the many separate individual relationships hypothesized for construct validity, it represented a simplified model that aimed to test the overall relationship of these variables and the factor structure. The model hypothesized that the same underlying factors that influence the various independent measures (e.g., WeeFIM, SDQ, WRAT) also influence the related HOQ domain scores. If this were the case (i.e., the model demonstrated a good fit), then it would suggest that the HOQ domain scores are, in fact, representative indicator variables for the proposed factors. In this way, the model was a test of construct validity. The results of the SEM, in fact, suggested a very good fit of the hypothesized model with all relative fit indices greater than 0.95 and a relatively small chi-squared statistic, although the p-value was 0.04. This provided added evidence of the construct validity of the HOQ. One concern, however, was that there was not good evidence for discriminant validity between the factors.[194] That is, the correlation between the Cognitive factor and the other two factors was rather high (greater than 0.60 in each case). This could be seen as weakness in the claim of construct validity for the HOQ, given the dual criteria of convergent and discriminant validity described by Campbell and Fiske.[195]

With a sample size of 80, this SEM also suffered from the limitation of a small sample size. This model had 24 unknown parameters, so, applying the 4:1 criteria for

sample size cited earlier, a sample of 96 subjects would have been preferable.[189, 192] However, despite this limitation and the expected effect that this might have on the fit indices, the model demonstrated excellent fit (based on relative indices).

7.3.3.3 Combining the Social-Emotional Domain

The decision to combine the Social and Emotional domains was based primarily on the desire to improve the reliability estimates of the subscore. However, this decision was made only after considering its conceptual implications. It was felt that these domains were conceptually similar and this, therefore, justified the combination of these domains into a single domain. This was tested further with the Validity testing by comparing correlations between the original Social and Emotional domains with the independent tests used for construct validity testing. While there were some differences between these sets of correlations, overall they were fairly similar, with a mean difference of 0.12. As well, the implications of the combined Social-Emotional domain on the proposed factor structure of the HOQ was tested. As hypothesized, the model which included the original Social and Emotional domains (**Model D**) had fit indices which were virtually identical to the model containing the combined Social-Emotional domain (**Model A**). These analyses provided further verification that the decision to combine these domains was conceptually sound and justified.

7.4 Summary

The HOQ was subjected to various forms of validity testing, including convergent construct validity, extreme group comparison, and confirmatory factor analysis. In each of these, the instrument performed largely as expected. While recognizing that the process of establishing construct validity is always an on-going one, these results suggest that the HOQ does appear to be a good measure of the health status of children with hydrocephalus.

8. Interpretation of Instrument Scores

8.1 Methods for Establishing Interpretability

As discussed in **Section 2.4**, there are many different ways of attributing interpretability to the numerical score of a health status instrument. Three general types of methods were investigated for the HOQ: the more traditional comparison of the numerical scores to a global health rating, a newer method described by the author in an earlier work (see **Section 2.4.3**),[1] and a data-driven effect size approach.[75, 76] The most important use of the HOQ will ultimately be as a research tool and, therefore, it is these type of comparisons for which it is most important to establish interpretability. To be more explicit, based on the taxonomy recently provided by Beaton et al., the following axes would be of interest:[196]

Who:	group differences
Which:	between-group
What:	observed change

8.1.1 Global Rating Comparison

During validity testing, global ratings of health were provided by the parents and the attending surgeons. The surgeons were surveyed just after having assessed the child and their family in a routine hospital clinic visit. They were asked to rate the child's health on a 6-point adjectival scale with the following descriptors:

Very severely impaired
Severely impaired
Moderately impaired
Mildly impaired
Very mildly impaired
Not at all impaired

Within each of the 6 response categories the median HOQ Total Score was calculated, along with the range of values. The distribution of HOQ Total Scores within each category was demonstrated graphically with a boxplot. As well, because the parents completed a second global rating assessment with the mailed survey (as described as part of the reliability testing in **Chapter 6**), a test-retest reliability was calculated for the global rating scale using ICC.

8.1.2 Comparison to the Health Utilities Index – 2 (HUI-2)

The HOQ Total Scores were also compared to the HUI-2 utility scores to assess

for a significant correlation. Based on the presence of a high correlation, a linear regression model was performed to allow for the transformation of HOQ Total Scores into utility scores. Two models were tested: a simpler model using the HOQ Total Score as the sole independent variable and a more complex model using the HOQ individual domain scores as the independent variables.

8.1.3 Effect Size Approach

This approach was briefly described in **Section 2.4.1.2.7**. The effect size is determined by taking the mean difference between two comparison groups and dividing this by the baseline standard deviation.[75, 76] The magnitude of the differences between the groups can then be gauged to the following benchmark effect size values: 0.20 = small difference, 0.50 = moderate difference, 0.80 = large difference.[75, 76] Conversely, beginning with these benchmark values and knowing the standard deviation of the HOQ, one can then calculate the mean difference in HOQ scores that would correspond to small, moderate, and large differences.

8.2 Results of Interpretability Testing

8.2.1 Global Rating Comparison

The overall correlation between global rating scores of the parents and surgeons to the HOQ Total Score was only moderate (Pearson correlation = 0.58 and 0.57, respectively). The test-retest ICC for the parents' global rating scale was 0.73 (0.55-0.84). The boxplots for the HOQ score distribution across the 6 global rating categories are shown below. The dark lines represent the median values, the boxes represent the interquartile distances, and the end lines represent the range of values:

FIGURE 8.1. Distribution of parents' scores

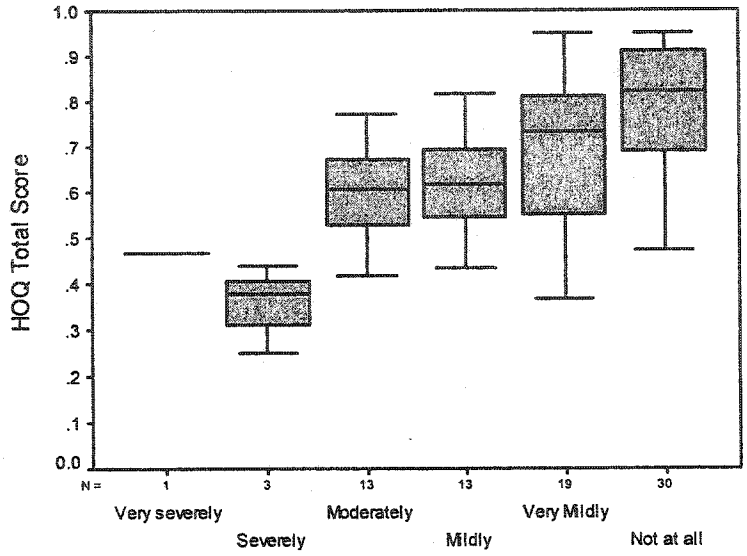
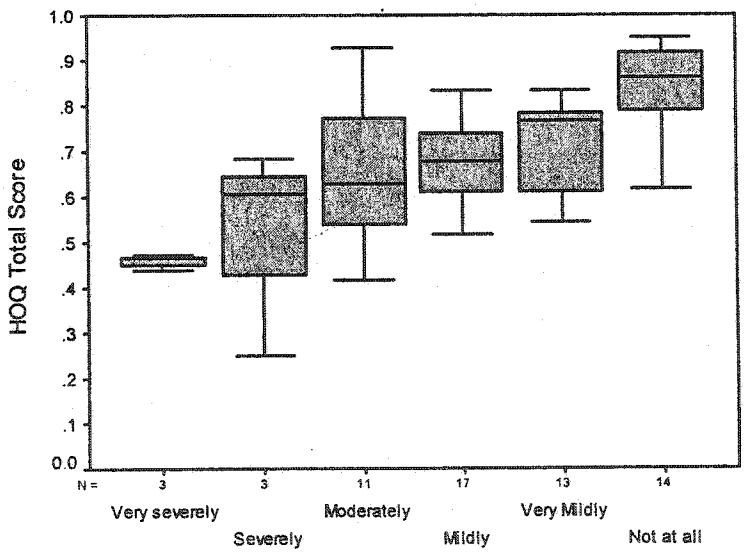


FIGURE 8.2. Distribution of surgeons' scores

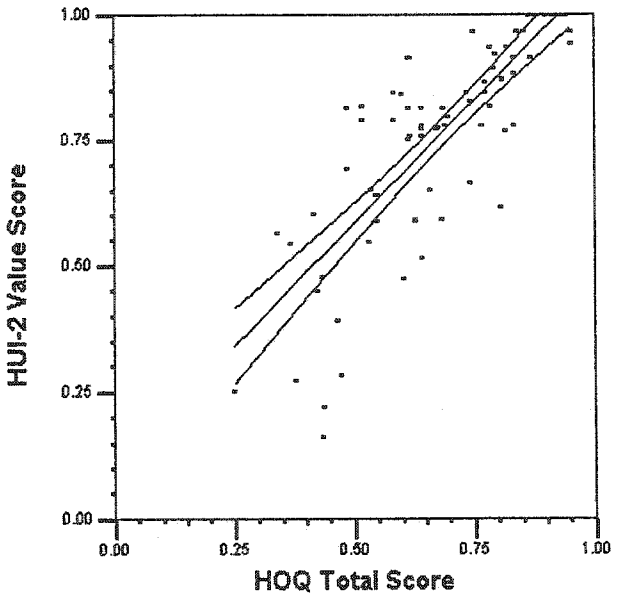


Because the correlation between these scores was rather poor and because there was substantial overlap of HOQ Scores between categories, it was felt that this method did not adequately address the issue of instrument interpretability.

8.2.2 Comparison to HUI-2

The correlation of the HUI-2 utility score and the HOQ Total Score was very high (0.81) and examination of a scatterplot suggested a strong linear relationship. Therefore, a linear regression analysis was performed. The results of this are demonstrated in the graph below:

FIGURE 8.3. HOQ and HUI-2 linear regression



The scatter plot shows the linear regression line of best fit, along with the mean 95% confidence band ($R^2 = 0.66$). The regression equation was:

HUI-2 Utility Score = 0.98 * (HOQ Total Score) + 0.10

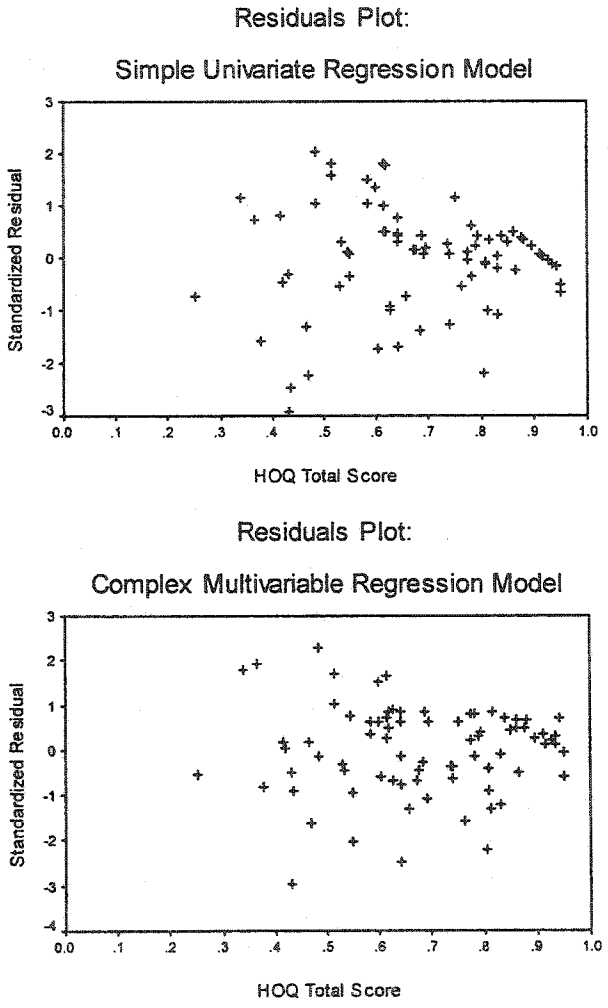
For comparison, a slightly more complex regression model was also tested by using the individual HOQ domain scores as the dependent variables, instead of using just the HOQ Total Score. This was more similar to the approach used by Nichol et al..[83] The following regression equation was determined:

HUI-2 Utility Score = 0.65 * (HOQ Physical Score)
+ 0.20 * (HOQ Social-Emotional Score)
+ 0.03 * (HOQ Cognitive Score) + 0.15

All parameters were significantly differently than zero, except for the β coefficient for

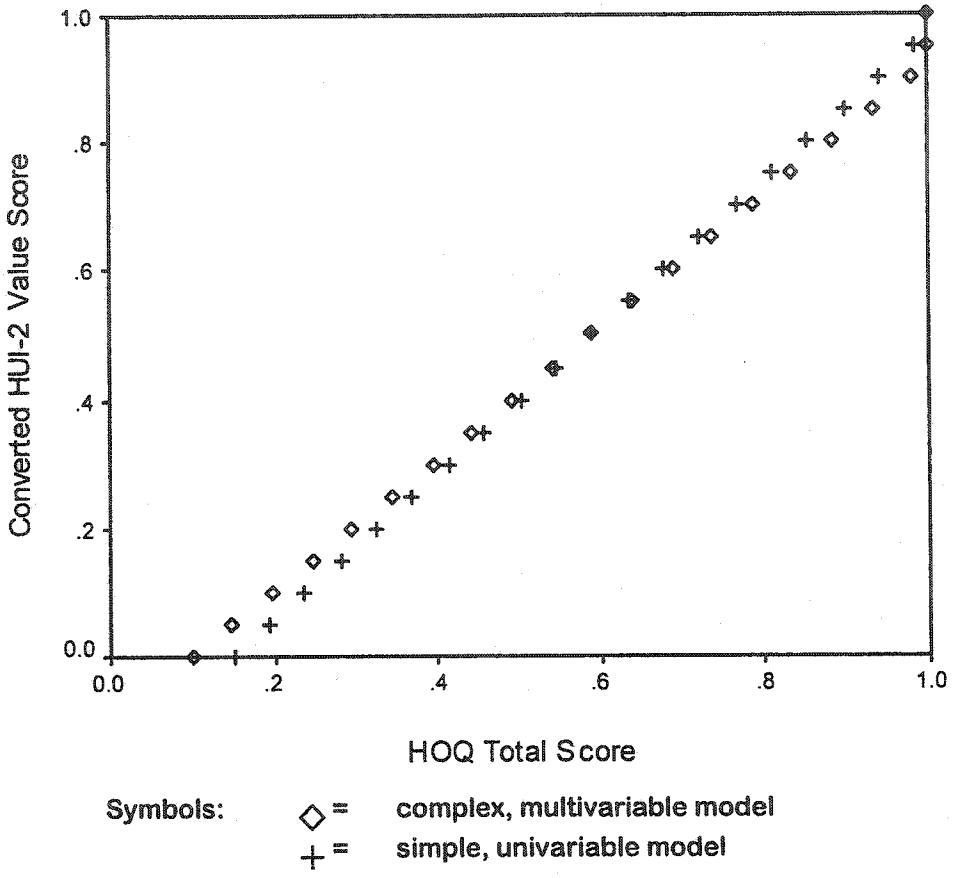
Cognitive Score (p=0.62). The R^2 of this model was 0.80, suggesting a better explanation of the variability than the simpler model. In order to examine this further graphically, plots of the standardized residuals of both models were compared and were found to be rather similar, with neither model providing a clearly better residual plot:

FIGURE 8.4 and 8.5. Residual plots of regression models



Furthermore, a comparison of the results of the two regression equations was plotted in comparison (for the complex model, it was artificially assumed that each of the HOQ domain scores was equivalent to the HOQ Total Score):

FIGURE 8.6. Comparison of regression model scores



Graphically, it appeared that both models provided rather similar HUI-2 estimates. A quadratic model was also tested. However, the R^2 was 0.67 – virtually identical to the simple univariable linear model.

8.2.3 Effect Size Approach

The standard deviations of the HOQ scores of the entire sample used in the validity and reliability studies were calculated. They are presented in the table below, along with the estimates of small, moderate, and large difference scores based on the effect size benchmarks of 0.20, 0.50, and 0.80, respectively:

TABLE 8.1. HOQ effect sizes

HOQ Domain	Mean Score	Standard Deviation	Small Difference	Moderate Difference	Large Difference
Physical	0.71	0.25	0.05	0.13	0.20
Social-Emotional	0.72	0.14	0.03	0.07	0.11
Cognitive	0.57	0.26	0.05	0.13	0.21
Total	0.68	0.17	0.03	0.09	0.14
Parental Concern	0.45	0.28	0.06	0.14	0.22

8.3 Discussion

As has been discussed earlier in this work (see **Section 2.4**), attributing interpretability to a new health status instrument is a difficult undertaking. One general type of approach has been to determine the minimum important difference (MID). The MID, as well as its many limitations, have also been discussed in detail in **Section 2.4.1**. Because of these limitations, it was decided that the MID approach would not be used to establish the interpretability of the HOQ. Instead, an approach that would provide some clinically relevant meaning to the HOQ scoring system was employed. Three methods were investigated, but one showed itself to be clearly inadequate. Attempting to compare the results of the parents' and surgeons' global ratings to the HOQ scores did not provide a successful means of interpreting the HOQ scores. Although there was a moderate correlation between the scoring systems, there was also a large degree of overlap in the categories. That is, the range of HOQ Total Scores for children who were globally rated as having their health impaired "Not at all" overlapped substantially with those children rated as having "Moderately" impaired health. A possible reason for this failure was the relatively modest reliability of the global rating scale (ICC of 0.73). This would attenuate any stronger relationship between the global ratings and the HOQ. As well, it is likely

that surgeon assessment of global rating is simply not a very accurate means of assessing a child's health status. While this may be adequate for more observable aspects of health, such as physical health, it is probably much more difficult to assess emotional and social health and provide an overall assessment which includes all these different facets of health. Although global ratings have been used successfully by others to determine the MID,[66] they did not prove useful in this case.

A second, newer method appeared to be more successful. A very good linear correlation was demonstrated between the HOQ and the HUI-2 utility scores. A linear regression model which accounted for a good deal of HUI-2 variability showed that the HOQ Total Score could be easily converted to a HUI-2 utility score, essentially by adding a value of 0.10. The more complex model involved using the individual domain scores in the regression equation. Although this had a higher R^2 value, the performance of both equations across the range of HOQ values was very similar, suggesting that the complex model added little extra information. Although this method proved successful statistically, it can still be argued that it does not further the cause of interpretability. That is, some clinicians and patients may not have any greater appreciation for a HOQ Total Score of 0.70, even if told that it equates to a utility score 0.79. It is true that conceptualizing the meaning of a utility score is not easy. However, one could counter-argue that it is just as difficult to conceptualize (at least consistently conceptualize) the meaning of "Almost the same, hardly any better at all", as used by others in establishing interpretability.[66] The added advantage of the utility score conversion is that it opens up the HOQ to a much wider variety of uses. For example, these values could be used in cost-utility analyses or for comparing the disease burden between diverse medical conditions. Based on the sample of hydrocephalic children accrued for the reliability/validity study, the mean HOQ Total Score was 0.68, which converts to a mean utility score of 0.77. This could be compared to some other utility scores obtained in the literature:

0.87..... teen-agers who were extremely low-birth weight[197]
0.86..... pediatric long-term survivors of liver transplant[198]
0.85..... pediatric survivors of Hodgkins's disease[199]
0.68..... adult stroke survivors[200]
0.58..... adults with Alzheimer's disease[200]

A brief look at this data gives some idea of the disease burden carried by children with hydrocephalus compared to these other common, chronic conditions.

A potential criticism of this method might arise from the high correlation observed between the simple HUI-2 and the longer HOQ. One might argue that it would be easier to administer the HUI-2 straight off, and simply ignore the HOQ entirely. However, while the correlations of the overall scores are high, the HOQ provides different and more detailed information about the health status of the child with hydrocephalus. For example, the correlation of individual HOQ domain scores with the sub-scores of HUI-2 was not uniformly high, suggesting that it is recording different information. The HOQ scores for Physical,

Social-Emotional, and Cognitive health can not all be replaced by the HUI-2. Another potential unique advantage of the HOQ is revealed by examining Figure 8.3. It appeared that the HUI-2 suffered from a ceiling effect in this population with a number of subjects scoring 1.0. The HUI-2 was incapable of discriminating amongst these subjects. The HOQ, however, did not demonstrate a ceiling effect for these high-scoring subjects.

The third method used to establish interpretability was the data-driven, effect size method. This used the external benchmark values of 0.20, 0.50, and 0.80 to represent effect size changes that were small, moderate, and large, respectively, as described by Cohen.[75, 76] Based on the standard deviations calculated from the sample used in this study, the magnitude of HOQ scores that would correspond to a "small" difference were in the range of 0.03 to 0.06. For a "moderate" difference, the values ranged from 0.07 to 0.14 and for a "large" difference, the values ranged from 0.11 to 0.22. Using these values, the differences calculated during the testing of extreme group comparisons (see **Section 7.2.4**) can be re-examined. For example, the difference in mean HOQ Cognitive score between patients with and without shunt infection was 0.17 – a "moderate" to "large" difference. The difference in HOQ Total Score between patients with and without epilepsy was 0.22 – a "large" difference. As has already been discussed in **Section 2.4.1.2**, this data-driven approach is one of many approaches that can be used. Although simple, some might be intuitively troubled by this approach because it is distribution-based and seemingly bypasses the use of any clinically relevant anchors. However, Samsa et al. argue that the original benchmark values of 0.20, 0.50, and 0.80 proposed by Cohen came out of an extensive empirical examination of various biological and psychological measures (e.g., the difference in heights of growing adolescents, the differences in individual IQ's, etc.).[75, 76] Perhaps, more importantly, they argue that these values have since proven to be quite consistent with what most clinicians would think intuitively about certain disease conditions. For example, they calculated the effect sizes for several disease conditions as follows, based on the SF-36, compared to a control population:[75]

- urinary tract infection	0.09
- depression	0.23
- angina	0.40
- osteoarthritis	0.52
- rheumatoid arthritis	0.76

One could reasonably argue that interpreting these values in the context of Cohen's benchmarks is very reasonable. As well, Angst et al., studying patients with osteoarthritis, compared the patient's global rating of their own change in health to the effect sizes for separate health status measures, e.g., the SF-36.[82] They defined the MID as the difference in the change scores between patients who rated themselves "slightly better/worse" and those who rated themselves "equal" after a 3-month period. The SF-36 effect size that corresponded to the clinically-anchored MID was 0.26, providing some further empirical validation of Cohen's original proposal.

If one accepts this use of effect sizes for establishing interpretability, this lends itself

quite easily to sample size calculation for studies in which the HOQ is an outcome measure. For example, with the knowledge that the standard deviation of the HOQ Total Score is 0.17, then one could design a comparative study that would aim to find a "moderate" difference of 0.09 between two groups. Assuming alpha=0.05 and a power of 90%, one would require just under 60 subjects per group for such a study.

8.4 Summary

Three different methods for helping to interpret the HOQ scores were investigated. While the comparison of the global ratings did not prove successful, the other two methods did. By providing a simple equation that converts the HOQ score into a utility score, the potential uses of the HOQ have been expanded, while also adding another way of interpreting the degree of health impairment associated with a given score. As well, by using the effect-size method, a range of values for each of the HOQ domain scores have been provided to allow for some basic interpretation of the differences seen between comparison groups.

9. Conclusion and Future Directions

This work began with the recognition that the currently available methods of measuring the health status of children with hydrocephalus were inadequate for various reasons. The goal was, therefore, to develop a new instrument that would specifically address these shortcomings. A detailed work has been presented in which a new health status outcome measurement instrument, called the Hydrocephalus Outcome Questionnaire (HOQ), has been developed. This is a simple and relatively easy-to-use questionnaire which is completed by the parents of children with hydrocephalus.

The HOQ was shown to be reliable and provided good initial evidence of validity. As well, different methods of establishing interpretability were also assessed. Overall, the HOQ appears to have met the criteria initially outlined for successful use as an outcome measure in clinical studies.

The current work represents merely the beginning of what will be a larger research effort into the health status of children with hydrocephalus. The directions of future research endeavours can be broadly classified as those concentrating on the use of the HOQ in different clinical studies and those that further explore the psychometric properties of the HOQ.

9.1 Future Clinical Studies

The main future clinical study that will be carried out with the HOQ will be a large cross-sectional study of children with hydrocephalus. This will involve administering the HOQ to a very large sample and simultaneously collecting numerous other data on these patients. These data will be used as independent variables in the final analyses and will include, for example, age, age at diagnosis, etiology, number of shunt revisions, number of shunt infections, frequency of seizures, type of shunt or ventriculostomy, ventricular size on imaging, and others. The sample will need to provide good representation of all the major etiologies of hydrocephalus and be large enough to allow for statistical analyses involving multiple independent variables. The goals of this study will be to provide an overall, comprehensive assessment of the current health status of children with hydrocephalus and to try to find associations between the independent variables and the HOQ. Given the methodological limitations of a cross-sectional study, these will be viewed as exploratory analyses, but will provide the basis for further, more specific studies in the future.

One particular analysis that will be especially interesting from this large study will be an examination of the effect of etiology on health status. As has been discussed earlier

in this work, hydrocephalus can be caused by a number of different underlying diseases. These vary widely and include brain hemorrhage, cerebral infection, brain tumours, head injury, myelomeningocele, aqueductal stenosis, and several others. It is quite possible that responses to the HOQ will differ amongst these different etiology groups. While the effect of etiology was explored in **Section 7.2.5** using a multivariable ANOVA, this did not appear to contribute significantly to the variation in the HOQ scores. However, this analysis was limited by a relatively small sample size. This was particularly a concern for the analysis of etiology because there were a number of etiology categories and there were relatively few representatives within each category. Therefore, a more useful analysis will require a much larger sample size. With such a sample size, differences in overall HOQ score and subscores amongst the different etiologies might become statistically evident. It will be interesting to compare, for example, the HOQ scores for children who developed hydrocephalus at or near birth (e.g., due to myelomeningocele or neonatal intraventricular hemorrhage) and those who developed hydrocephalus later in life (e.g., due to head injury or infection).

Another future clinical study will involve using the HOQ in other centres outside of Toronto. This will assess how generalizable the HOQ is and how easy it will be to transfer the protocol and scoring system to outside institutions. This will also allow collaborative research efforts among various centres. Initially, this will be limited to centres in English-speaking countries. However, if there is sufficient interest, it is possible for the HOQ to be translated into other languages. The process of translating health status instruments is fairly demanding and should follow a methodologically sound protocol.[201] If such a translation process is undertaken successfully, this will open up the uses of the HOQ to involve cross-cultural and international comparisons of the health status.

9.2 Exploration of Measurement Properties

The HOQ was developed primarily as a discriminative instrument for use in group-comparison studies. During this work, these were the properties which were assessed and quantified. However, it is possible that the instrument possesses other properties which have not been assessed. An example of this is the ability of the HOQ to function as an evaluative instrument. For reasons outlined at the outset of this work, it was not with this intent that the HOQ was developed. Nonetheless, it would be, at least, of academic interest to assess whether the HOQ does possess such properties. In the context of hydrocephalus, the clinical scenarios in which the evaluative properties can be tested are rather limited. It would require identifying a subgroup of children with hydrocephalus who have a relatively stable period of health impairment for which some form of intervention could be performed. After this intervention, it would be expected that some children will improve while others would not. One of the few such subgroups in the hydrocephalic population are those children who present with long history of

cognitive or social-emotional difficulties and are then found to have previously undiagnosed, chronic hydrocephalus. The one other group is those children with known hydrocephalus who have similar long-standing deficits due to inadequate or improperly functioning CSF shunt systems. In both groups, surgical intervention may help improve some, but likely not all, of the children. It is possible for the HOQ to be administered to these children prior to their surgical intervention and then in post-operative follow-up. The changes in the HOQ scores of patients who were deemed to have improved, using some external standard, e.g., overall parental assessment, could be compared to those children who did not improve. From this the responsiveness of the instrument could be calculated. Using the taxonomy of Beaton et al., the axes of interest would be:[196]

Who:	group differences
Which:	over time
What:	estimated change

Another area in which to explore the measurement properties of the HOQ is to better assess the child-version of the HOQ. Some preliminary results have been presented in this work (see **Section 6.2.4**). The results of the reliability testing in this small sample of children were promising. However, a much larger sample would need to be assessed in order to truly know how well the HOQ will perform when completed by older, cognitively-capable children. This study would also allow for a more detailed comparison of results between child and parent responses.

9.3 Conclusions

This work has established the groundwork necessary for future studies into the health status of children with hydrocephalus. The HOQ has demonstrated very good psychometric properties and future studies will be aimed at using it in wider, multicentre settings. Further work will also be done to future explore the measurement properties of the HOQ, which may then further widen the potential uses of the instrument.

10. Bibliography

1. Kulkarni AV. Development of a disease-specific health status measurement for children with hydrocephalus (M.Sc. thesis). Department of Clinical Epidemiology and Biostatistics. Hamilton: McMaster University, 1999.

2. Elwood J. Major central nervous system malformations in Northern Ireland, 1969 to 1973. Develop Med Child Neurol 1976; 18:512-520.

3. Milhorat T. Pediatric Neurosurgery. Philadelphia: Davis, 1978.

4. DelBigio M. Epidemiology and direct economic impact of hydrocephalus: a community based study. Can J Neurol Sci 1998; 25:123-126.

5. Stein S, Schut L. Hydrocephalus in myelomeningocele. Child's Brain 1979; 5:413-419.

6. Hoffman H, Smith M. The use of shunting devices for cerebrospinal fluid in Canada. Can J Neurol Sci 1986; 13:81-87.

7. Bondurant C, Jimenez D. Epidemiology of cerebrospinal fluid shunting. Pediatr Neurosurg 1995; 23:254-258.

8. Eisenberg H, McComb J, Lorenzo A. Cerebrospinal fluid overproduction and hydrocephalus associated with choroid plexus papilloma. J Neurosurg 1974; 40:381-5.

9. Milhorat T. Hydrocephalus: Pathophysiology and Clinical Features. In: Wilkins R, Rengachary S, eds. Neurosurgery. New York: McGraw-Hill, 1996.

10. Chumas PD, Armstrong DC, Drake JM, et al. Tonsillar herniation: the rule rather than the exception after lumboperitoneal shunting in the pediatric population. J

Neurosurg 1993; 78:568-73.

11. Chumas PD, Kulkarni AV, Drake JM, Hoffman HJ, Humphreys RP, Rutka JT. Lumboperitoneal shunting: a retrospective study in the pediatric population. Neurosurgery 1993; 32:376-83; discussion 383.

12. Tuli S, Drake JM. Multiple shunt failures: an analysis of relevant features. Childs Nerv Syst 1999; 15:79.

13. Tuli S, Drake J, Lawless J, Wigg M, Lamberti-Pasculli M. Risk factors for repeated cerebrospinal shunt failures in pediatric patients with hydrocephalus. J Neurosurg 2000; 92:31-8.

14. Drake JM, Kestle JR, Tuli S. CSF shunts 50 years on--past, present and future. Childs Nerv Syst 2000; 16:800-4.

15. Liechty E, Gilmor R, Bryson C, Bull M. Outcome of high-risk neonates with ventriculomegaly. Deveop Med Child Neurol 1983; 25:162-168.

16. Maixner W, Morgan M, Besser M, Johnston I. Ventricular volume in infantile hydrocephalus and its relationship to intracranial pressure and cerebrospinal fluid clearance before and after treatment. Pediatr Neurosurg 1990-91; 16:191-196.

17. Renier D, Sainte-Rose C, Pierre-Kahn A, Hirsch J. Prenatal hydrocephalus: outcome and prognosis. Childs Nerv Syst 1988; 4:213-222.

18. Shankaran S, Slovis T, Bedard M, Poland R. Sonographic classification of intracranial hemorrhage. A prognostic indicator of mortality, morbidity, and short-term neurologic outcome. J Pediatr 1982; 100:469-475.

19. Shankaran S, Koepke T, Woldt E, et al. Outcome after posthemorrhagic ventriculomegaly in comparison with mild heorrhage without ventriculomegaly. J Pediatr 1989; 114:109-114.

20. Casey A, Kimmings E, Kleinlugtebeld A, Taylor W, Harkness W, Hayward R.

The long-term outlook for hydrocephalus in childhood. A ten-year cohort study of 155 patients. Pediatr Neurosurg 1997; 27:63-70.

21. Hoppe-Hirsch E, Laroussinie F, Brunet L, et al. Late outcome of the surgical treatment of hydrocephalus. Childs Nerv Syst 1998; 14:97-9.

22. Lumenta C, Skotarczak U. Long-term follow-up in 233 patients with congenital hydrocephalus. Childs Nerv Sys 1995; 11:173-175.

23. Iskandar B, Tubbs S, Mapstone T, Grabb P, Bartolucci A, Oakes W. Death in shunted hydrocephalic children in the 1990s. Pediatr Neurosurg 1998; 28:173-6.

24. Drake J, Kestle J, Milner R, et al. Randomized trial of cerebrospinal fluid shunt valve design in pediatric hydrocephalus. Neurosurgery 1998; 43:294-303.

25. Kulkarni AV, Rabin D, Lamberti-Pasculli M, Drake JM. Repeat cerebrospinal fluid shunt infection in children. Pediatr Neurosurg 2001; 35:66-71.

26. Kulkarni AV, Drake JM, Lamberti-Pasculli M. Cerebrospinal fluid shunt infection: a prospective study of risk factors. J Neurosurg 2001; 94:195-201.

27. Donders J, Canady A, Rourke B. Psychometric intelligence after infantile hydrocephalus. Childs Nerv Syst 1990; 6:148-154.

28. Fernell E, Hagberg G, Hagberg B. Infantile hydrocephalus epidemiology: an indicator of enhanced survival. Arch Dis Child 1994; 70:F123-F128.

29. Barnes M, Dennis M. Reading in children and adolescents after early onset hydrocephalus and in normally developing age peers: phonological analysis, word recognition, word comprehension, and passage comprehension skill. J Pediatr Psychol 1992; 17:445-65.

30. Barnes M, Dennis M. Discourse after early-onset hydrocephalus: core deficits in children of average intelligence. Brain Language 1998; 61:309-34.

31. Dennis M, Barnes M. Oral discourse after early-onset hydrocephalus: linguistic ambiguity, figurative language, speech acts, and script-based inferences. J Pediatr Psychol 1993; 18:639-52.

32. Dennis M, Jacennik B, Barnes M. The content of narrative discourse in children and adolescents after early-onset hydrocephalus and in normally developing age peers. Brain Language 1994; 46:129-65.

33. Fletcher J, Landry S, Bohan T, et al. Effects of intraventricular hemorrhage and hydrocephalus on the long-term neurobehavioral development of preterm very-low-birthweight infants. Dev Med Child Neurol 1997; 39:596-606.

34. Thompson N, Fletcher J, Chapieski L, Landry S, Miner M, Bixby J. Cognitive and motor abilities in preschool hydrocephalics. J Clin Exp Neuropsychol 1991; 13:245-58.

35. Bourgeois S, Sainte-Rose C, Cinalli G, et al. Epilepsy in children with shunted hydrocephalus. J Neurosurg 1999; 90:274-81.

36. Noetzel M, Blake J. Seizures in children with congenital hydrocephalus: long-term outcome. Neurology 1992; 42:1277-81.

37. Piatt J, Carlson C. Hydrocephalus and epilepsy: an actuarial analysis. Neurosurgery 1996; 39:722-728.

38. Donders J, Rourke B, Canady A. Emotional adjustment of children with hydrocephalus and of their parents. J Child Neurol 1992; 7:375-80.

39. Fernell E, Gillberg C, vonWendt L. Behavioural problems in children with infantile hydrocephalus. Dev Med Child Neurol 1991; 33:388-95.

40. Fletcher J, Brookshire B, Landry S, et al. Behavioral adjustment of children with hydrocephalus: relationships with etiology, neurological, and family status. J Pediatr Psychol 1995; 20:109-125.

41. Stellman-Ward G, Bannister C, Lewis M, Shaw J. The incidence of chronic headache in children with shunted hydrocephalus. Europ J Pediatr Surg 1997; 1:12-4.

42. James H, Nowak T. Clinical course and diagnosis of migraine headaches in hydrocephalic children. Pediatr Neurosurg 1991; 17:310-6.

43. Dennis M, Fitz C, Netley C, et al. The intelligence of hydrocephalic children. Arch Neurol 1981; 38:607-615.

44. Query J, Reichelt C, Christoferson L. Living with chronic illness: a retrospective study of patients shunted for hydrocephalus and their families. Develop Med Child Neurol 1990; 32:119-128.

45. Kirshner B, Guyatt G. A methodological framework for assessing health indices. J Chron Dis 1985; 38:27-36.

46. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40:171-178.

47. Tugwell P, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982; 9:758-762.

48. Allen M, Yen W. Introduction to Measurement Theory. Belmont, Calif.: Wadsworth Inc., 1979.

49. Nunnally J. Introduction to Psychological Measurement. New York: McGraw-Hill, 1970.

50. Streiner D, Norman G. Health Measurement Scales: A Practical Guide to Their Development and Use. New York: Oxford University Press, 1995:231.

51. Cronbach L, Rajaratnam N, Gleser G. Theory of generalizability: A liberazation of reliability theory. Br J Stat Psychol 1963; 16:137-163.

52. Cronbach L, Gleser G, Nanda H, Rajaratnam N. The Dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

53. Burns K. Beyond classical reliability: using generalizability theory to assess dependability. Res Nurs Health 1998; 21:83-90.

54. Brennan R. Elements of Generalizability Theory. Iowa City, Iowa: ACT Publications, 1983.

55. Shavelson R, Webb N. Generalizability theory (in press). In: Kempf-Leonard K, ed. Encyclopedia of Social Measurement. San Diego, Calif.: Academic Press, 2002.

56. Juniper E, Guyat G, Epstein R, Ferrie P, Jaeschke R, Hiller T. Evaluation of impairment of health-related quality of life in asthma: development of a questionnaire for use in clinical trials. Thorax 1992; 47:76-83.

57. Marks G, Dunn S, Woodcock A. A scale for the measurement of quality of life in adults with asthma. J Clin Epidemiol 1992; 45:461-472.

58. Varni J, Katz E, Seid M, Quiggins D, Friedman-Bender A. The pediatric cancer quality of life inventory-32 (PCQL-32): I. Reliability and validity. Cancer 1998; 82:1184-96.

59. Parkin P, Kirpalani H, Rosenbaum P, et al. Development of a health-related quality of life instrument for use in children with spina bifida. Qual Life Res 1997; 6:123-132.

60. Miller G. The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol Bull 1956; 63:81-97.

61. Guyatt G, Townsend M, Berman L, Keller J. A comparison of Likert and visual analogue scales for measuring change in function. J Chron Dis 1987; 40:1129-1133.

62. Nishisato N, Torii Y. Effects of categorizing continuous normal distributions on the product-moment correlation. Jap Psychol Res 1970; 13:45-49.

63. Hughes S, Dodder R. Alcohol consumption indices: format comparisons. J Stud Alcohol 1988; 49:100-103.

64. Berwick D, Budman S, Damico-White J, Feldstein M, Klerman G. Assessment of psychological morbidity in primary care: explorations wth the General Health Questionnaire. J Chron Dis 1987; 40 (Suppl 1):71S-79S.

65. Guyatt G. Making sense of quality-of-life data. Med Care 2000; 38 (Suppl. II):175-179.

66. Jaeschke R, Singer J, Guyatts G. Measurement of health status. Ascertaining the minimal clinically important difference. Cont Clin Trials 1989; 10:407-415.

67. Wells G, Beaton D, Shea B, et al. Minimal clinically important differences: review of methods. J Rheumatol 2001; 28:406-412.

68. Aseltine R, Carlson K, Fowler F, Barry M. Comparing prospective and retrospective measures of treatment outcomes. Med Care 1995; 33 (suppl):AS67-AS76.

69. Redelmeier D, Guyat G, Goldstein R. Assessing the minimal important difference in symptoms: a comparison of two techniques. J Clin Epidemiol 1996; 49:1215-1219.

70. Bellamy N, Carette S, Ford P, et al. Osteoarthritis antirheumatic drug trials. III. Setting the delta for clinical trials - results of a consensus development (Delphi) exercise. J Rheumatol 1992; 19:451-457.

71. vanWalraven C, Mahon J, Moher D, Bohm C, Laupacis A. Surveying physicians to determine the minimal important difference: implications for sample-size calculation. J Clin Epidemiol 1999; 52:717-723.

72. Todd K, Funk J. The minimum clinically important difference in physician-assigned visual analog pain scores. Acad Emerg Med 1996; 3:142-146.

73. Stratford P, Riddle D, Binkley J, Spadoni G, Westaway M, Padfield B. Using the neck disability index to make decisions concerning individual patients. Physiother Can 1999; Spring:107-112.

74. Wyrwich K, Tierney W, Wolinsky F. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999; 52:861-873.

75. Samsa G, Edelman D, Rothman M, Williams R, Lipscomb J, Matchar D. Determing clinically important differences in health status measures. A general approach with illustration to the Health Utilities Index Mark II. Pharmacoeconomics 1999; 15:141-155.

76. Cohen J. Statistical power analysis for the behavioral sciences, 2nd ed.. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1988.

77. Felson D, Anderson J, Lange M, Wells G, LaValley M. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998; 41:1564-1570.

78. Riddle D, Stratford P, Binkley J. Sensitivity to change of the Roland-Morris back pain questionnaire. Phys Ther 1998; 78:1197-1207.

79. Kirwan J. Minimum clinically important difference: the crock of gold at the end of the rainbow? J Rheumatol 2001; 28:439-444.

80. Lydick E. Approaches to the interpretation of quality-of-life scales. Med Care 2000; 38 (Suppl. II):180-183.

81. Testa M. Interpretation of quality-of-life outcomes. Issues that affect magnitude and meaning. Med Care 2000; 38 (Suppl. II):166-174.

82. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients wiht osteoarthritis of the lower extremities. Arthritis Care Res 2001; 45:384-391.

83. Nichol M, Sengupta N, Globe D. Evaluating quality-adjusted life years: estimation of the health utility index (HUI2) from the SF-36. Med Decis Making 2001; 21:105-112.

84. Last J. A Dictionary of Epidemiology, Third Edition. New York: Oxford University Press, Inc., 1995.

85. Lenert L, Kaplan R. Validity and interpretation of preference-based measures of health-related quality of life. Med Care 2000; 38 (Suppl. II):138-150.

86. Macran S, Kind P. "Death" and the valuation of health-related quality of life. Med Care 2001; 39:217-227.

87. Kaplan R, Feeny D, Revicki D. Methods for assessing relative importance in preference based outcome measures. Qual Life Res 1993; 2:467-475.

88. Feeny D. A utility approach to the assessment of health-related quality of life. Med Care 2000; 38 (Suppl. II):151-154.

89. Boyle M, Torrance G, Sinclair J, et al. Economic evaluation of neonatal intensive care of very low birth weight infants. New Engl J Med 1983; 308:1330.

90. Torrance G, Boyle M, Horwood S. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res 1982; 30.

91. Feeny D, Furlong W, Barr R, Torrance G, Rosenbaum P, Weitzman S. A comprehensive multiattribute system for classifying the health status of survivors of childhood cancer. J Clin Oncol 1992; 10:923-928.

92. Gemke R, Bonsel G. Reliability and validity of a comprehensive health status measure in a heterogeneous population of children admitted to intensive care. J Clin Epidemiol 1996; 49:327-333.

93. Saigal S, Feeny D, Furlong W, Rosenbaum P, Burrows E, Torrance G. Comparison of the health-related quality of life of extremely low birth weight children and a reference group of children at age eight years. J Pediatr 1994; 125:418-425.

94. Revicki D, Kaplan R. Relationship between psychometric and utility-based approaches to the measurement of health-related quality of life. Qual Life Res 1993; 2:477-487.

95. Hoyle R. Structural Equation Modeling. Concepts, Issues, and Applications. Thousand Oaks, CA: Sage Publications, 1995.

96. McDonald R, Ho M. Principles and practice in reporting structural equation analyses. Psychol Methods 2002; 7:64-82.

97. Schumacker R, Lomax R. A Beginner's Guide to Structural Equation Modeling. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc., 1996.

98. Bryant F, Yarnold P, Michelson E. Statistical methodology: VIII. Using confirmatory factor analysis (CFA) in emergency medicine research. Acad Emerg Med 1999; 6:54-66.

99. Bryman A. Social Research Methods. Oxford, UK.: Oxford University Press, 2001.

100. Crabtree B, Miller W. Doing Qualitative Research, 2nd ed. Thousand Oaks: Sage Publications Inc., 1999.

101. Dilorio C, Hockenberry-Eaton M, Maibach E, Rivero T. Focus groups: an interview method for nursing research. J Neurosci Nurs 1994; 26:175-180.

102. Morgan D. Planning Focus Groups. Thousand Oaks, CA.: Sage, 1998.

103. Carey M. Comment: Concerns in the analysis of focus group data. Qualit Health Res 1995; 5:487-495.

104. Tang K, Davis A. Critical factors in the determination of focus group size. Family Practice 1995; 12:474-475.

105. Livingstone S, Lunt P. Talk on Television: Audience Participation and Public Debate. London: Routledge, 1994.

106. Asbury J. Overview of focus group research. Qualit Health Res 1995; 5:414-420.

107. Henderson N. A practical approach to analyzing and reporting focus groups studies: lessons from qualitative market research. Qualit Health Res 1995; 5:463-477.

108. Mangione T. Mail Surveys: Improving the Quality. Thousand Oaks, CA: Sage Publications, 1995.

109. Yammarino F, Skinner S, Childers T. Understanding mail survey response behavior. Public Opinion Quaterly 1991; 55:613-639.

110. Anema M, Brown B. Increasing survey responses using the Total Design Method. J Cont Edu Nurs 1995; 26:109-114.

111. McCrohan K, Lowe L. A cost/benefit approach to postage used on mail questionnaires. J Marketing 1981; 45:130-133.

112. WHO. Health Promotion : A Discussion Document. Copenhagen: WHO, 1984.

113. Saracci R. The World Health Organisation needs to reconsider its definition of health. BMJ 1997; 314:1409-10.

114. Susser M. Ethical components in the definition of health. Int J Health Serv 1974; 4:539-548.

115. World Health Organization. International Classification of Impairments, Disabilities, and Handicaps. Geneva: WHO, 1980.

116. Thuriaux M. The ICIDH: evolution, status, and prospects. Disabil Rehabil 1995; 17:112-118.

117. Brandsma JW, Lakerveld-Heyl K, Van Ravensberg CD, Heerkens YF. Reflection on the definition of impairment and disability as defined by the World Health Organization. Disabil Rehabil 1995; 17:119-27.

118. World Health Organization. International Classification of Functioning, Disability and Health. Copenhagen: WHO, 2001.

119. Keller M. Toward a definition of health. ANS Adv Nurs Sci 1981; 4:43-64.

120. Patrick D, Bush J, Chen M. Toward an operational definition of health. J Health Soc Behav 1973; 14:6-23.

121. Bergner M. Measurement of health status. Med Care 1985; 23:696-704.

122. Eisen M, Ware J, Donald C, Brook R. Measuring components of children's health status. Med Care 1979; 17:902-921.

123. Cadman D, Goldsmith C, Bashim P. Values, preferences and decisions in the care of children with developmental disabilities. J Dev Behav Pediatr 1984; 5:60-64.

124. Pal D. Quality of life assessment in children: a review of conceptual and methodological issues in multidimensional health status measures. J Epidemiol Community Health 1996; 50:391-396.

125. Pantell R, Lewis C. Measuring the impact of medical care on children. J Chron

Dis 1987; 40 (Suppl 1):99S-108S.

126. Juniper E, Guyatts G, Feeny D, Griffith L, Ferrie P. Minimum skills required by children to complete health-related quality of life instruments for asthma: comparison of measurement properties. Eur Respir J 1997; 10:2285-2294.

127. Harter S. The Perceived Competence Scale for Children. Child Development 1982; 53:87-97.

128. Juniper E, Howland W, Roberts N, Thompson A, King D. Measuring quality of life in children with rhinoconjunctivitis. J Allergy Clin Immunol 1998; 101:163-170.

129. Wolman C, Basco D. Factors influencing self-esteem and self-consciousness in adolescents with spina bifida. J Adolesc Health 1994; 15:543-548.

130. Stein R, Jessop D. Functional Status II(R). A measure of child health status. Med Care 1990; 28:1041-1055.

131. Starfield B, Bergner M, Ensminger M, et al. Adolescent health status measurement: development of the Child Health and Illness Profile. Pediatrics 1993; 91:430-5.

132. Spencer N, Coe C. The development and validation of a measure of parent-reported child health and morbidity: the Warwick Child Health and Morbidity Profile. Child: Care, Health & Development 1996; 22:367-79.

133. Langeveld J, Koot H, Loonen M, Hazebroek-Kampschreur A, Passchier J. A quality of life instrument for adolescents with chronic headache. Cephalalgia 1996; 16:183-196.

134. Budd K, Workman D, Lemsky C, Quick D. The Children's Headache Assessment Scale (CHAS): factor structure and psychometric properties. J Behav Med 1994; 17:159-179.

135. McCormick M, Charney E, Stemmler M. Assessing the impact of a child with spina bifida on the family. Dev Med Child Neurol 1986; 28:53-61.

136. Hoare P, Russell M. The quality of life of children with chronic epilepsy and their families: preliminary findings with a new assessment measure. Dev Med Child Neurol 1995; 37:689-96.

137. Kalucy M, Bower C, Stanley F. School-aged children with spina bifida in Western Australia - parental perspectives on functional outcome. Dev Med Child Neurol 1996; 38:325-34.

138. Goodman R. A modified version of the Rutter parent questionnaire including extra items on children's strengths: a research note. J Child Psychol Psychiat 1994; 35:1483-1494.

139. Juniper E, Guyatts G, Streiner D, King D. Clinical impact versus factor analysis for quality of life questionnaire construction. J Clin Epidemiol 1997; 50:233-238.

140. Hyland M, Finnis S, Irvine S. A scale for assessing quality of life in adult asthma sufferers. J Psychosomatic Res 1991; 35:99-110.

141. Marx R, Bombardier C, Hogg-Johnson S, Wright J. How should importance and severity ratings be combined for item reduction in the development of health status instruments? J Clin Epidemiol 1999; 52:193-197.

142. Marx R, Bombardier C, Hogg-Johnson S, Wright J. Clinimetric and psychometric strategies for development of a health measurement scale. J Clin Epidemiol 1999; 52:105-111.

143. Msall ME, DiGaudio K, Rogers BT, et al. The Functional Independence Measure for Children (WeeFIM). Conceptual basis and pilot use in children with developmental disabilities. Clin Pediatr 1994; 33:421-30.

144. Stein R, Riessman C. The development of an impact-on-family scale: preliminary findings. Med Care 1980; 18:465-471.

145. Goodman R, Meltzer H, Bailey V. The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. Europ Child Adol Psychiatry 1998; 7:125-30.

146. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987; 6:441-448.

147. Walter S, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998; 17:101-110.

148. Deyo R, Diehr P, Patrick D. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Cont Clin Trials 1991; 12:142S-158S.

149. Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420-428.

150. Green S, Salkind N, Akey T. Using SPSS for Windows: Analyzing and Understanding Data. Upper Saddle River, N.J.: Prentice Hall, 2000.

151. Hays R, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life. Qual Life Res 1993; 2:441-449.

152. Birmaher B, Khetarpal S, Brent D, et al. The Screen for Child Anxiety Related Emotional Disorders (SCARED): scale construction and psychometric characteristics. J Am Acad Child Adolesc Psychiatry 1997; 36:545-53.

153. Daltroy L, Liang M, Fossel A, Goldberg M. The POSNA pediatric musculoskeletal functional health questionnaire: report on reliability, validity, and sensitivity to change. J Pediatr Orthop 1998; 18:561-571.

154. Doherty E, Yanni G, Conroy R, Bresnihan B. A comparison of child and parent ratings of disability and pain in juvenile chronic arthritis. J Rheumatol 1993; 20:1563-1566.

155. Duffy C, Arsenault L, Duffy K. Level of agreement between parents and children in rating dysfunction in juvenile rheumatoid arthritis and juvenile spondyloarthritides. J Rheumatol 1993; 20:2134-2139.

156. Eiser C, Havermans T, Craft A, Kernahan J. Development of a measure to assess the perceived illness experience after treatment for cancer. Arch Dis Child 1995; 72:302-307.

157. Erling A, Wiklund I, Albertsson-Wikland K. Prepubertal children with short stature have a different perception of their well-being and stature than their parents. Qual Life Res 1994; 3:425-429.

158. Glaser A, Davies K, Walker D, Brazier D. Influence of proxy respondents and mode of administration on health status assessment following central nervous system tumours in childhood. Qual Life Res 1997; 6:43-53.

159. Guyatt G, Junniper E, Griffith L, Feeny D, Ferrie P. Children and adult perceptions of childhood asthma. Pediatrics 1997; 99:165-168.

160. Hays R, Vickrey B, Hermann B, et al. Agreement between self reports and proxy reports of quality of life in epilepsy patients. Qual Life Res 1995; 4:159-168.

161. Reid G, Gilbert C, McGrath P. The Pain Coping Questionnaire: preliminary validation. Pain 1998; 76:83-96.

162. Rothman M, Hedrick S, Bulcroft K, Hickam D, Rubinstein L. The validity of proxy-generated scores as measures of patient health status. Med Care 1991; 29:115-124.

163. Sprangers M, Aaronson N. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992; 45:743-760.

164. Theunissen N, Vogel T, Koopman H, et al. The proxy problem: child report versus parent report in health-related quality of life research. Qual Life Res 1998;

7:387-397.

165. Bellman M, Paley C. Parents underestimate children's pain. (letter). BMJ 1993; 307:1563.

166. Vogels T, Verrips G, Verloove-Vanhorick S, et al. Measuring health-related quality of life in children: the development of the TACQOL parent form. Qual Life Res 1998; 7:457-65.

167. Weissman M, Orvaschel H, Padian N. Children's symptom and social functioning self-report scales. Comparison of mothers' and children's reports. J Nerv Ment Dis 1980; 168:736-740.

168. Goodman R. The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. J Child Psychol Psychiatry 1999; 40:791-9.

169. Boyle M, Torrance G. Developing multi-attribute health indexes. Med Care 1984; 22:1045-1057.

170. Boyle M, Furlong W, Feeny D, Torrance G, Hatcher J. Reliability of the Health Utilities Index - Mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995; 4:249-257.

171. Trudel J, Rivard M, Dobkin P, Leclerc J, Robaey P. Psychometric properties of the Health Utilities Index Mark 2 system in paediatric oncology patients. Qual Life Res 1998; 7:421-432.

172. Furlong W, Feeny D, GW GT, Barr R. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. Ann Med 2001; 33:375-384.

173. Jessop D, Riessman C, Stein R. Chronic childhood illness and maternal mental health. J Dev Behav Pediatr 1988; 9:147-156.

174. Juniper E, Guyatt G, Feeny D, Ferrie P, Griffith L, Townsend M. Measuring quality of life in the parents of children with asthma. Qual Life Res 1996; 5:27-34.

175. Czyzewski D, Mariotto M, Bartholomew L, Lecompte S, Sockrider M. Measurement of quality of well being in a child and adolescent cystic fibrosis population. Med Care 1994; 32:965-972.

176. Wilkinson G. Wide Range Achievement Test 3. Adminstration Manual. Wilmington, Delaware: Wide Range, Inc., 1993.

177. Arbuckle J, Wothke W. AMOS 4.0 Users Guide. Chicago, IL: SPSS inc., 1999.

178. Tucker L, Lewis C. The reliability coefficient for maximum likelihood factor analysis. Psychometrika 1973; 38:1-10.

179. Bollen K. Sample size and Betler and Bonett's nonnormed fit index. Psychometrika 1986; 51:375-377.

180. Bentler P, Chou C. Practical issues in structural modeling. Soc Meth Res 1987; 16:78-117.

181. Marsh H, Balla J, McDonald R. Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. Psychol Bull 1988; 103:391-410.

182. Tanguma J. Effects of sample size on the distribution of selected fit indices: a graphical approach. Edu Psychol Measurement 2001; 61:759-776.

183. Streiner D. Figuring out factors: the use and misuse of factor analysis. Can J Psychiatry 1994; 39:135-140.

184. Zwick W, Velicer W. Comparison of five rules for determining the number of components to retain. Psychol Bull 1986; 99:432-442.

185. Rubio D, Berg-Weger M, Tebb S. Using structural equation modeling to test for

multidimensionality. Structural Equation Model 2001; 8:613-626.

186. Laurence KM, Evans RC, Weeks RD, Thomas MD, Frazer AK, Tew BJ. The reliability of prediction of outcome in spina bifida. Dev Med Child Neurol Suppl 1976:150-6.

187. Hunt GM, Oakeshott P, Kerry S. Link between the CSF shunt and achievement in adults with spina bifida. J Neurol Neurosurg Psychiatry 1999; 67:591-5.

188. Cate IM, Kennedy C, Stevenson J. Disability and quality of life in spina bifida and hydrocephalus. Dev Med Child Neurol 2002; 44:317-22.

189. Hoyle R. Evaluating measurement models in clinical research: covariance structure analysis of latent variable models of self-conception. J Consult Clin Psychol 1991; 59:67-76.

190. Martinez M, Marshall J, Sechrest L. Factor analysis and the search for objectivity. Am J Epidemiol 1998; 148:17-19.

191. Foelker G, Shewchuk R, Niederehe G. Confirmatory factor analysis of the short form Beck Depression Inventory in elderly community samples. J Clin Psychol 1987; 43:111-118.

192. Tanaka J. How big is big enough? Sample size and goodness of fit in structural equation models with latent variables. Child Development 1987; 58:134-146.

193. Hu L, Bentler P. Evaluating model fit. In: Hoyle R, ed. Structural Equation Modeling. Concepts, Issues, and Applications. Thousand Oaks, CA: Sage Publications, 1995.

194. Cole D. Utility of confirmatory factor analysis in test validation research. J Consult Clin Psychol 1987; 55:584-594.

195. Campbell D, Fiske D. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959; 56:81-105.

196. Beaton D, Bombardier C, Katz J, Wright J. A taxonomy for responsiveness. J Clin Epidemiol 2001; 54:1204-1217.

197. Saigal S, Feeny D, Rosenbaum P, Furlong W, Burrows E, Stoskopf B. Self-perceived health status and health-related quality of life of extremely low-birth-weight infants at adolescence. JAMA 1996; 276:453-459.

198. Midgley DE, Bradlee TA, Donohoe C, Kent KP, Alonso EM. Health-related quality of life in long-term survivors of pediatric liver transplantation. Liver Transpl 2000; 6:333-9.

199. Van Schaik CS, Barr RD, Depauw S, Furlong W, Feeny D. Assessment of health status and health-related quality of life in survivors of Hodgkin's disease in childhood. Int J Cancer Suppl 1999; 12:32-8.

200. Mittmann N, Trakas K, Risebrough N, Liu B. Utility scores for chronic conditions in a community-dwelling population. Pharmacoeconomics 1999; 15:369-376.

201. Ware J, Keller S, Gandek B, Brazier J, Sullivan M. Evaluating translations of health status questionnaires. Methods from the IQOLA project. International Quality of Life Assessment. Intl J Technol Assess Health Care 1995; 11:525-551.

11. Appendices

Appendix A – Information Sheet for Item Generation from Expert Sources

Development of a Disease-Specific Health Status Measurement for Children with Hydrocephalus

Investigators: James M. Drake, M.D. and Abhaya V. Kulkarni, M.D.

Objective:
Currently there is no good outcome instrument that measures the health status of the hydrocephalic child from the *family perspective*. The objective of this study is to develop just such an instrument, that will ultimately take the form of a 40-50 item questionnaire that *parents themselves* can quite easily complete. This questionnaire will ask about the items that the parents would consider the most important and relevant in determining *their child's health status and well-being*. The development of this instrument will follow the rigid methodological protocol that has been previously well established.

The ultimate goal is to develop a questionnaire that will provide us with a scientific measure of how well the child is doing, taken from the parents' perspective. This would provide useful information in long-term follow-up studies and in clinical trials. This will allow us to define successful treatment in terms other than simply shunt failure, for example.

Why we need your help
We are currently in the phase of questionnaire development called *item generation*. The goal of this is to get a comprehensive list of all possible items that might be relevant to a parent's perspective of their child's health status and well being.

For this phase of the questionnaire development, we need your help, as a member of the health-care profession who deals extensively with children with hydrocephalus and their families. **We are asking for you to think about the *issues that parents of children with hydrocephalus consider important.*** Any issue or item that you think might even be remotely important to parents should be considered. For example, these may be issues related to cognitive, physical, or social problems or to the shunt, for example. These may be issues that are frequently brought up by parents during clinic visits or on the ward.

Once we have collected all possible items, the list will eventually be reduced to just 40-50 items during the next phase, called *item selection.*

We kindly ask that you list down all the items you can think of on the attached sheet.

They need only be listed in point form and you may use the back of the sheet if necessary. As well, please take the time to complete the basic information about yourself at the top of the attached page.

Thank you very much for your co-operation and time. If you have any questions, **please feel free to contact Dr. Kulkarni**

Name:
(optional)_____

Position:

_____
(e.g., surgeon, clinic nurse, ward nurse, etc.)

Years of Experience in Dealing with Children with Hydrocephalus: _____

Please list all items that you think parents would consider important in determining their child's overall health-status and well being in each of the following areas:

1. Physical Health

2. Social Health

3. Emotional Health

4. Cognitive Health

5. Level of Independence and Self-Care

6. Pain

7. Communication

8. School Performance

9. Future Development/Prognosis

10. Any Other Areas (please feel free to use back of pages if necessary)

Appendix B – Sample Consent Form for Focus Group Participation

CLINICAL INFORMATION FORM
(parents – focus group)

Title of Research Project: **Development of a Health Status Outcome Measure for Pediatric Hydrocephalus**

Investigators: **Dr. James M. Drake**
Dr. Abhaya V. Kulkarni

Name:

Date of Birth:

HSC #:

Purpose of Research

The purpose of this study is to develop a scale for children with hydrocephalus in order to assess how well they are doing. Specifically, this will take the form of a questionnaire that can be administered to parents. However, in order to develop a questionnaire that works, we need input from parents, like yourselves.

Description of Research

Participation in the study would involve you participating in a small group discussion with other parents who also have children with hydrocephalus. The group discussion will be informal and will attempt to address many of the issues that are of concern to parents of children with hydrocephalus. For example, there will be discussion about your child's physical, social, and emotional well-being. This group session will last between 2 to 3 hours.

The discussion will be informal and casual. You can feel free to say whatever you like, but you also have the option of not partaking in certain parts of the discussion directly. This is entirely up to you.

These sessions will be videotaped but this will only be for the use of the research investigator. These will not be used in any sort of public forum. The videotaping is needed to ensure that we do not miss any of the important points that will be brought up during the group discussion.

Part of this research will also involve the investigators reviewing the medical chart of your child.

Potential Harms, Discomforts, or Inconvenience

If you want to be part of the group discussion, it may take about 2 to 3 hours of your time. Potentially, some of the questions may be of a personal nature. If some of the questions upset you, you

don't have to answer them. This is the benefit of being in a group setting: no one person is put on the spot to answer any questions. As well, you are free to leave at anytime during the session.

Potential Benefits

There is no direct benefit to you by participating in this study. However, by speaking with many parents, we hope to be able to understand the things that are most important in making your children's lives as healthy as possible. Hopefully, we can then use this information to help treat children in the best possible way.

Alternatives

As there is no therapeutic implication to this study, there is no direct alternative to participation.

Confidentiality

Confidentiality will be respected and no information that discloses the identity of the subject will be released or published without consent unless required by the law. For your information, the research consent form will be inserted in the patient health record.

Participation

Participation in this study is entirely voluntary. If you choose not to participate, you and your child will continue to have access to quality care at HSC. If you choose to participate, you can withdraw at anytime, without in any way affecting your access to quality care at HSC.

Consent

I acknowledge that the research procedures described above have been explained to me and that any questions that I have asked have been answered to my satisfaction. I have been informed of the alternatives to participation in this study, including the right not to participate and the right to withdraw without compromising the quality of medical care at the Hospital for Sick Children for my child and for other members of my family. As well, the potential harms and discomforts have been explained to me and I also understand the benefits (if any) of participating in the research study. I know that I may ask now, or in the future, any questions I have about the study or the research procedures. I have been assured that records relating to my child and my child's care will be kept confidential and that no information will be released or printed that would disclose personal identity without my permission unless required by law.

I hereby consent to participate.

Appendix C. Preliminary List of Items Generated During First Phase of Research.

- Physical Health

To what extent is your child able to walk normally without assistance?
How well does your child sleep at night?
How much does the presence of a shunt limit your child's physical abilities?
To what extent is your child free of physical disability?
How well does your child cope with his/her physical disabilities?
To what extent is your child able to enjoy sports?
To what extent does fatigue limit your child's activities?
How well is your child able to run?
How well is your child able to write?
How well is your child able to tie his/her shoelaces?
To what extent does your child experience falls?
How difficult is it for your child to walk up stairs?
To what extent do seizures limit your child's activities?
How concerned is your child about having further seizures?

- Social Health

How easy is it for your child to make new friends?
To what extent does your child feel welcome in the homes of others?
To what extent is your child able to maintain friends?
How well does your child get along with his/her siblings?
How well does your child get along with you?
To what extent is your child able to visit his/her friends?
To what extent is your child able to go on family trips?
To what extent is your child treated as an equal by his/her peers?
To what extent does your child feel part of the family?
To what extent does your child participate in after-school activities?

- Emotional Health

To what extent does your child feel embarrassed about his/her illness?
How much does your child express his/her emotions?
How would you rate your child's self-esteem?
How happy is your child?
How strong an emotional bound is there between you and your child?
How bothered is your child about visible shunt tubing or scars?
To what extent does your child feel that he/she is not "normal"?

- Cognitive Health

How much does the presence of a shunt effect your child's cognitive abilities?
To what extent is your child free of cognitive disability?
How well does your child cope with his/her cognitive disabilities?
To what extent is your child motivated to learn?
How well does your child remember things?
How well is your child able to concentrate on a task, e.g., schoolwork?
To what extent does your child appear to be slower at learning?
How well is your child able to read?

- Level of Independence and Self-Care

To what extent does your child need help eating?
To what extent does your child need help dressing?
To what extent does your child need help going to the washroom?
To what extent does your child need help bathing?
How well does your child know his/her daily routine?

- Pain

To what extent does child suffer from headaches?
To what extent is your child free of pain?
To what extent does your child suffer from tummy pains?
To what extent does your child feel pain along the shunt tubing?

- Communication

To what extent is your child able to see normally?
To what extent is your child able to hear normally?
To what extent is your child able to speak normally?
How well is your child able to express him/herself?
To what extent does your child need repeat instructions to perform tasks?

- School Performance

How proud are you of your child's academic achievements?
How well does your child pay attention in school?
To what extent is your child able to attend a regular school?
How much does your child enjoy school?
To what extent is your child treated as normal by his/her teachers?
To what extent does your child need lessons repeated?
How frustrated does your child feel about schoolwork?

- Future Development/Prognosis

How proud are you of your child's progress?
To what extent does your child worry about his/her future?
To what extent do you worry about your child's future?
To what extent do you feel that your child will be able to choose his/her own career path?
To what extent do you feel that your child will be able to live independently?
How concerned are you about the ability of your child to have children of her own?
How concerned are you about your child's progression through the school system?

- Other

To what extent do you feel that your child is accepted and valued by society?
How much time does your child spend within the medical system?
How much of a burden does your child's illness place on your marriage/relationship/partner?
How much does your child's illness interfere with your nurturing of the other children?
How much does your child's illness interfere with normal family activities?
To what extent do hospital visits disrupt your child's life?
To what extent are you concerned about the medications your child takes?
To what extent are you concerned about the need for future shunt surgery?
To what extent are you concerned about shunt infections?
To what extent are you concerned about future shunt malfunctions?
How concerned are you that your child will outgrow his/her shunt?

Appendix D - Item Reduction Questionnaire (selected items)

My child:	How TRUE is the statement for your child?					How IMPORTANT is this to your child's well-being?				
	Not at all	A little	Somewhat	Quite a bit	Very	Not at all	A little	Somewhat	Quite a bit	Very
does well at school.....	O	O	O	O	O	O	O	O	O	O
enjoys school.....	O	O	O	O	O	O	O	O	O	O
is not treated as a "normal" student by teachers.....	O	O	O	O	O	O	O	O	O	O
needs special educational assistance, e.g. tutor.....	O	O	O	O	O	O	O	O	O	O
is slower at doing homework.....	O	O	O	O	O	O	O	O	O	O
has difficulty in participating in extra-curricular activities.....	O	O	O	O	O	O	O	O	O	O
is easily frustrated.....	O	O	O	O	O	O	O	O	O	O
is frustrated by schoolwork.....	O	O	O	O	O	O	O	O	O	O
is frustrated by homework.....	O	O	O	O	O	O	O	O	O	O
is frustrated by social encounters with peers.....	O	O	O	O	O	O	O	O	O	O
has difficulty learning at school....	O	O	O	O	O	O	O	O	O	O
undergoes physical therapy.....	O	O	O	O	O	O	O	O	O	O
undergoes occupational therapy...	O	O	O	O	O	O	O	O	O	O
has difficulty hearing.....	O	O	O	O	O	O	O	O	O	O
has difficulty speaking.....	O	O	O	O	O	O	O	O	O	O
forgots his/her daily routines.....	O	O	O	O	O	O	O	O	O	O
copes well with his/her disabilities..	O	O	O	O	O	O	O	O	O	O
needs help eating food.....	O	O	O	O	O	O	O	O	O	O
needs help dressing.....	O	O	O	O	O	O	O	O	O	O
takes a long time to get ready for school.....	O	O	O	O	O	O	O	O	O	O
needs help going to the washroom at home.....	O	O	O	O	O	O	O	O	O	O
needs help going to the washroom at school.....	O	O	O	O	O	O	O	O	O	O
needs help bathing.....	O	O	O	O	O	O	O	O	O	O
has difficulty verbally expressing his/her wants or needs.....	O	O	O	O	O	O	O	O	O	O

Appendix E – Cover Letter for Item Reduction Questionnaire

Dr. James Drake and I work in the Division of Neurosurgery at the Hospital for Sick Children. We are asking for assistance in a hydrocephalus research project we are conducting. The purpose of this study is to develop a way of measuring how well children with hydrocephalus are doing (outcome). This will be done by asking their families questions about their child's health. In order to develop a questionnaire that works, we need to know what questions to ask and how to ask them. For this we need input from the children's families.

What we are asking from you:

1. I have enclosed a questionnaire. This questionnaire addresses many of the issues that are of concern to parents of children with hydrocephalus. It asks about your child's physical, social, and emotional well-being. It also asks about your concerns as parents. This questionnaire takes about 60 minutes to complete.
I have enclosed 2 copies of the questionnaire, so that both mother and father, if available, can complete it separately.

2. Please read and sign the enclosed Clinical Information Form (consent form).

3. After you have completed the questionnaires and the Clinical Information Form, please return all of them together in the envelope provided. I have already paid for the postage.

In order to answer any questions you might have, I will be calling your home shortly after you receive this package. In addition, if you have any questions at all, please feel free to contact me. You can page me, day or night, at
will do my best to return your page as soon as possible.

Thank you very much for your time. It is greatly appreciated and will prove to be very useful as we try to improve our treatment of children with hydrocephalus.

Kindest regards,

Abhaya V. Kulkarni, M.D., M.Sc., Division of Neurosurgery

Appendix F – 62-Selected Items for Reliability and Validity Testing

HYDROCEPHALUS OUTCOME QUESTIONNAIRE

(to be completed by the parents)

Child's name: Completed by: O mother O father

Hospital ID #:

Date completed:

Date of birth:

Please FILL IN THE CIRCLE that best represents your answer to HOW TRUE the following STATEMENTS are about YOUR CHILD over the last 4 WEEKS:

MY CHILD:	Not at all true	A little true	Somewhat true	Quite a bit true	Very true
1. needs help dressing	O	O	O	O	O
2. needs help going to the washroom	O	O	O	O	O
3. has poor vision	O	O	O	O	O
4. has difficulty walking	O	O	O	O	O
5. needs a wheelchair	O	O	O	O	O
6. has difficulty participating in sports	O	O	O	O	O
7. has difficulty with hand-writing	O	O	O	O	O
8. has poor physical balance	O	O	O	O	O
9. has difficulty tying shoe-laces	O	O	O	O	O
10. gets tired easily	O	O	O	O	O
11. has difficulty speaking	O	O	O	O	O
12. suffers from headaches	O	O	O	O	O
13. has many seizures	O	O	O	O	O
14. needs help bathing	O	O	O	O	O
15. needs help eating food	O	O	O	O	O
16. has difficulty participating in extra-curricular activities	O	O	O	O	O
17. feels like he/she is being stared at in public	O	O	O	O	O
18. has difficulty separating from me	O	O	O	O	O
19. has many friends	O	O	O	O	O
20. is treated as an equal by his/her peers	O	O	O	O	O
21. is able to visit his/her friends	O	O	O	O	O
22. is solitary and keeps to him/herself	O	O	O	O	O
23. has difficulty recognizing the consequences of his/her actions	O	O	O	O	O

24. misses a lot of school due to illness	O	O	O	O	O
25. gets anxious in social situations	O	O	O	O	O
26. has difficulty getting along with his/her peers	O	O	O	O	O
27. is shy in public	O	O	O	O	O
28. has difficulty playing with his/her peers	O	O	O	O	O
29. is easily frustrated	O	O	O	O	O
30. copes well with his/her disabilities	O	O	O	O	O
31. has difficulty verbally expressing his/her feelings	O	O	O	O	O
32. feels confident about him/herself	O	O	O	O	O
33. often feels stressed	O	O	O	O	O

34. is often irritable	O	O	O	O	O
35. is unmotivated	O	O	O	O	O
36. is concerned about his/her physical appearance	O	O	O	O	O
37. often feels sad	O	O	O	O	O
38. worries about the future	O	O	O	O	O
39. is often restless	O	O	O	O	O
40. lacks self-confidence	O	O	O	O	O
41. over-reacts to other people's illnesses	O	O	O	O	O
42. has a poor concept of time	O	O	O	O	O
43. has difficulty with math	O	O	O	O	O
44. is well organized	O	O	O	O	O
45. has difficulty concentrating	O	O	O	O	O
46. is forgetful	O	O	O	O	O
47. has difficulty performing several tasks in a row	O	O	O	O	O
48. has difficulty reading	O	O	O	O	O
49. is a slow learner	O	O	O	O	O
50. needs instructions repeated	O	O	O	O	O
51. forgets his/her daily routines	O	O	O	O	O
52. has difficulty learning new tasks	O	O	O	O	O
53. has a short attention span	O	O	O	O	O

Appendix F - continued

Please FILL IN THE CIRCLE that best represents your answer to HOW TRUE the following STATEMENTS are about YOUR CONCERNS about YOUR CHILD:

I AM CONCERNED:	Not at all true	A little true	Somewhat true	Quite a bit true	Very true
1. about my child's ability to take care of a family in the future	O	O	O	O	O
2. about my child's ability to find a career	O	O	O	O	O
3. about my child's ability to live alone in the future	O	O	O	O	O
4. about the need for future shunt surgery	O	O	O	O	O
5. about my child's future education	O	O	O	O	O
6. about my child's ability to maintain friendships	O	O	O	O	O
7. about my child's ability to participate in sports in the future	O	O	O	O	O
8. that my child's shunt will become blocked	O	O	O	O	O
9. that my child's shunt will become infected	O	O	O	O	O

THANK YOU VERY MUCH.

Appendix G – Functional Independence Measure for Children[143]

	Total Assist. (<25%) 1	Max. Assist. (>25%) 2	Mod. Assist. (>50%) 3	Min. Assist. (>75%) 4	Super-vision 5	Mod. Indep. (device) 6	Full Indep. 7
1. Eating: Use of cup, spoon, & fork.....	O	O	O	O	O	O	O
2. Grooming: Tooth-brushing, hair-grooming, washing hands/face.....	O	O	O	O	O	O	O
3. Bathing: Washing, rinsing, drying the body from neck down.....	O	O	O	O	O	O	O
4. Dressing: upper Dressing/undressing above the waist, incl. fasteners/prosthetics.....	O	O	O	O	O	O	O
5. Dressing: lower Dressing/undressing below the waist, incl. fasteners/prosthetics.....	O	O	O	O	O	O	O
6. Toileting: Maintaining personal hygeine, adjusting clothes before/after toilet use.....	O	O	O	O	O	O	O
7. Bladder management: Maintaining urinary continency.....	O	O	O	O	O	O	O
8. Bowel management: Maintaining bowel continency.....	O	O	O	O	O	O	O
9. Transfers chair/wheelchair: Getting in/out of chair/wheelchair....	O	O	O	O	O	O	O
10. Transfers toilet: Getting on/off toilets.....	O	O	O	O	O	O	O
11. Transfers tub/shower: Getting in/out of tubs/showers.....	O	O	O	O	O	O	O
12. Walk/wheelchair/crawl: Most frequent mode of locomotion...	O	O	O	O	O	O	O
13. Stairs: Going up/down 12-14 indoor stairs...	O	O	O	O	O	O	O
14. Comprehension: Understanding of visual or auditory communication.....	O	O	O	O	O	O	O
15. Expression: Spoken/gesture language.....	O	O	O	O	O	O	O

Appendix G - continued

16. Social interaction:							
Sharing & taking turns with peers during play.....	O	O	O	O	O	O	O
17. Problem-solving:							
Carrying out sequence of self-correcting tasks (cleaning spill, telephone use)....	O	O	O	O	O	O	O
18. Memory:							
Orientation, routines (songs/games), recent special events (holidays/B-days).....	O	O	O	O	O	O	O

Appendix H – Strengths and Difficulties Questionnaire[145]

	Not True	Somewhat True	Certainly True
1. Considerate of other people's feelings.....	O	O	O
2. Restless, overactive, cannot stay still for long.....	O	O	O
3. Often complains of headaches, stomach-aches, or sickness.....	O	O	O
4. Shares readily with other children (treats, toys, pencils, etc.) ..	O	O	O
5. Often has temper tantrums or hot tempers.....	O	O	O
6. Rather solitary, tends to play alone.....	O	O	O
7. Generally obedient, usually does what adults request.....	O	O	O
8. Many worries, often seems worried.....	O	O	O
9. Helpful if someone is hurt, upset or feeling ill.....	O	O	O
10 .Constantly fidgeting or squirming.....	O	O	O
11. Has at least one good friend.....	O	O	O
12. Often fights with other children or bullies them.....	O	O	O
13. Often unhappy, down-hearted or tearful.....	O	O	O
14. Generally liked by other children.....	O	O	O
15. Easily distracted, concentration wanders.....	O	O	O
16. Nervous or clingy in new situations, easily loses confidence.	O	O	O
17. Kind to younger children.....	O	O	O
18. Often lies or cheats.....	O	O	O
19. Picked on or bullied by other children.....	O	O	O
20. Often volunteers to help others (parents, teachers, other children)..	O	O	O
21. Thinks things out before acting.....	O	O	O
22. Steals from home, school or elsewhere.....	O	O	O
23. Gets on better with adults that with other children.....	O	O	O
24. Many fears, easily scared.....	O	O	O
25. Sees task through to the end, good attention span.....	O	O	O

Appendix I – Health Utilities Index - 2[90]

Attribute	Level	Description
Sensation	1.....	Able to see, hear, and speak normally for age.
	2.....	Requires equipment to see or hear or speak.
	3.....	Sees, hears, or speaks with limitations even with equipment.
	4.....	Blind, deaf, or mute.
Mobility	1.....	Able to walk, bend, lift, jump, and run normally for age.
	2.....	Walks, bends, lifts, jumps, or runs with some limitations but does not require help.
	3.....	Requires mechanical equipment (such as canes, crutches, braces, or wheelchair) to walk or get around independently.
	4.....	Requires the help of another person to walk or get around and requires mechanical equipment as well.
	5.....	Unable to control or use arms and legs.
Emotion	1.....	Generally happy and free from worry.
	2.....	Occasionally fretful, angry, irritable, anxious, depressed, or suffering "night terrors".
	3.....	Often fretful, angry, irritable, anxious, depressed, or suffering from "night terrors".
	4.....	Almost always fretful, angry, irritable, anxious, or depressed.
	5.....	Extremely fretful, angry, irritable, anxious, or depressed, usually requiring hospitalization or psychiatric institutional care.
Cognition	1.....	Learns and remembers school work normally for age.
	2.....	Learns and remembers school work more slowly than classmates as judged by parents and/or teachers.
	3.....	Learns and remembers very slowly and usually requires special educational assistance.
	4.....	Unable to learn and remember.

Appendix I - continued

Self-care	1.....	Eats, bathes, dresses, and uses the toilet normally for age.
	2.....	Eats, bathes, dresses, and uses the toilet independently with difficulty.
	3.....	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently.
	4.....	Requires the help of another person to eat, bathe, dress, or use the toilet.
Pain	1.....	Free of pain discomfort.
	2.....	Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities.
	3.....	Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities.
	4.....	Frequent pain; frequent disruption of normal activities. Discomfort requires prescription narcotics for relief.
	5.....	Severe pain. Pain not relieved by drugs and constantly disrupts normal activities.

Appendix J – Impact-on-Family Scale[144]

	Strongly Disagree			Strongly Agree
Financial Burden:				
1. Additional income is needed in order to cover medical expenses.....	O	O	O	O
2. The illness is causing financial problems for the family.....	O	O	O	O
3. Time is lost from work because of hospital appointments.	O	O	O	O
4. I am cutting down on the hours I work to care for my child.	O	O	O	O

	Strongly Disagree			Strongly Agree
Familial/Social Impact:				
5. Our family gives up things because of my child's illness.	O	O	O	O
6. People in the neighbourhood treat us specially because of my child's illness.	O	O	O	O
7. We see family and friends less because of the illness.....	O	O	O	O
8. I don't have much time left over for other family members after caring for my child.....	O	O	O	O
9. We have little desire to go out because of my child's illness.	O	O	O	O
10. Because of the illness we are not able to travel out of the city.	O	O	O	O
11. Sometimes we have to change plans about going out at the last minute because of my child's state.	O	O	O	O
12. Sometimes I wonder if my child should be treated "specially" or the same as a normal child.....	O	O	O	O
13. I think about not having more children because of the illness.....	O	O	O	O

	Strongly Disagree			Strongly Agree
Personal Strain:				
14. Nobody understands the burden I carry.	O	O	O	O
15. Traveling to the hospital is a strain on me.	O	O	O	O
16. Sometimes I feel like we live on a roller-coaster: in crisis when my child is acutely ill, OK when things are stable.	O	O	O	O
17. It is hard to find a reliable person to take care of my child.....	O	O	O	O
18. I live from day to day and don't plan for the future.	O	O	O	O
19. Fatigue is a problem for me because of my child's illness.	O	O	O	O

Appendix J - continued

	Strongly Disagree			Strongly Agree
Mastery:				
20. Learning to manage my child's illness has made me feel better about myself.....	O	O	O	O
21. Because of what we have shared we are a closer family.....	O	O	O	O
22. My partner and I discuss my child's problems together.	O	O	O	O
23. We try to treat my child as if he/she were a normal child.	O	O	O	O
24. My relatives have been understanding and helpful with my child.....	O	O	O	O

Appendix K – Sample Consent Form for Reliability Testing

CLINICAL INFORMATION FORM
(parents – questionnaire)

Title of Research Project: **Development of a Health Status Outcome Measure for Pediatric Hydrocephalus**

Investigators: **Dr. James M. Drake**
Dr. Abhaya V. Kulkarni

Name:

Date of Birth:

HSC #:

Purpose of Research

The purpose of this study is to develop a scale for children with hydrocephalus in order to assess how well they are doing. Specifically, this will take the form of a questionnaire that can be administered to parents. However, in order to develop a questionnaire that works, we need the input of parents, like yourselves.

Description of Research

Participation in the study would involve you actually filling out a trial questionnaire. This questionnaire addresses many of the issues that are of concern to parents of children with hydrocephalus. It asks about your child's physical, social, and emotional well-being. This questionnaire takes about 20-30 minutes to complete. When possible, both parents will be asked to complete the questionnaire separately.

After you have completed this questionnaire, we will be mailing you another similar questionnaire in approximately 2 weeks time. We would ask that both parents, once again, complete the questionnaire separately. Once it is complete, we ask that you mail it back to us (a postage-paid envelope will be included).

Part of this research will also involve the investigators reviewing the medical chart of your child.

Potential Harms, Discomforts, or Inconvenience

If you want to be part of the study, it may take about 20-30 minutes of time commitment now and another 20-30 minutes at home to complete both sets of questionnaires. Potentially, some of the questions may be of a personal nature. If some of the questions upset you, you don't have to answer them and you can stop at any time.

Potential Benefits

There is no direct benefit to you by participating in this study. However, by speaking with many families, we hope to be able to understand the things that are most important in making your children's lives as happy as possible. Hopefully, we can then use this information to help treat children in the best possible way.

Alternatives

As there is no therapeutic implication to this study, there is no direct alternative to participation.

Confidentiality

Confidentiality will be respected and no information that discloses the identity of the subject will be released or published without consent unless required by the law. For your information, the research consent form will be inserted in the patient health record.

Participation

Participation in this study is entirely voluntary. If you choose not to participate, you and your child will continue to have access to quality care at HSC. If you choose to participate, you and your child can withdraw at anytime, without in any way affecting your access to quality care at HSC.

Consent

I acknowledge that the research procedures described above have been explained to me and that any questions that I have asked have been answered to my satisfaction. I have been informed of the alternatives to participation in this study, including the right not to participate and the right to withdraw without compromising the quality of medical care at the Hospital for Sick Children for my child and for other members of my family. As well, the potential harms and discomforts have been explained to me and I also understand the benefits (if any) of participating in the research study. I know that I may ask now, or in the future, any questions I have about the study or the research procedures. I have been assured that records relating to my child and my child's care will be kept confidential and that no information will be released or printed that would disclose personal identity without my permission unless required by law.

I hereby consent to participate.