SUCCESSES AND PITFALLS IN RUNNING A SMALL PROGRAM MMI

SUCCESSES AND PITFALLS IN RUNNING A SMALL PROGRAM MMI

By CATHY HUMPHREYS BASc, CLSt. dipl, CCLS

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science in Health Science Education

McMaster University MASTER OF SCIENCE IN HEALTH SCIENCE EDUCATION (2016) Hamilton Ontario

TITLE: Successes and Pitfalls in Running a Small Program MMI

AUTHOR: Cathy Humphreys, BASc. (Guelph) CLSt. dipl. (McMaster), CCLS

SUPERVISOR: Dr. Harold Reiter

NUMBER OF PAGES Pages: xi, 99

ABSTRACT

Purpose: Academic programs are faced with the important task of selecting health professional students who not only possess necessary cognitive abilities to perform their future job, but also have valuable personal/professional characteristics to draw upon in the provision of quality patient care. There is therefore "widespread agreement that it is desirable to broaden the scope of assessment beyond academic achievement" (Eva, Reiter, Trinh, Wasi, Rosenfeld, Norman, 2009, p.768) in candidate selection. The Child Life Studies Program at McMaster University used a 4-station multiple mini-interview (MMI) as part of the admission selection process. This study sought to determine the feasibility, degree of acceptability and reliability of this 4-station MMI design, and if there were any predictors of candidate's performance on the MMI.

Methods: A group of 35 applicants in 2014, and 40 applicants in 2015 screened through admission procedures participated in a 4-station MMI. Each station was 15 minutes in length with 5 minutes for scoring each candidate. Anonymous stakeholder surveys were used in 2015 to assess participant and interviewer's perceptions and acceptability of the MMI in applicant selection. Generalizability coefficients were calculated to determine reliability. In addition, candidate's experience in healthcare, professional work experience, experience with children with disabilities, and previous child life course work (or lack thereof) were analyzed through independent t-tests to report any relationship with candidate's performance on the MMI. A one-way ANOVA was also completed to report any relationship between candidate's undergraduate degree type and his/her MMI performance.

Results: The 4-station MMI was found to be feasible for a small program with a marked increase found in the number of candidates interviewed in a shorter period of time. It was also found to be highly acceptable among candidates and faculty. However, this studies' findings did not show statistically significant differences in MMI performance based on identified predictors, or undergraduate degree type. The reliability of the 4-station design in 2014 was G=0.718, however, was only G=0.089 in 2015, far lower than expected based on the prior year's G Coefficient. The potential pitfalls in running a small program MMI are discussed with specific suggestions and modifications provided to enhance reliability of candidate selections across professions.

ACKNOWLEDGEMENTS

I would like to express by heartfelt thanks to my thesis committee. To Dr. Harold Reiter, my thesis supervisor, thank you for your expertise, mentorship and continuous support throughout this journey. Your encouragement and insight have been invaluable. To Dr. Kelly Dore, and Dr. Geoff Norman, thank you for your support, guidance and helpful suggestions. I have learned so much from all of you.

To Molly Pottruff, thank you for inputting the data for the statistical analysis. I sincerely appreciate all of your time, patience and thoughtful suggestions.

To my husband Dave, thank you for your unwavering support, and ongoing encouragement throughout my graduate studies. Your willingness to take on extra responsibilities while I tried to navigate being a student, a mom to our young family, and work full-time has been overwhelming. You are without question my biggest cheerleader, and best friend. I love you more each day.

Most of all, thank you to my beautiful children Emma and Quinn. You've grown and developed so much since my graduate studies began. Thank you for your patience, your hugs and kisses, and for showing everyday what's most important in life. Our family is my biggest blessing, and you will both forever be my best accomplishment. Now, let's play!

Table of Contents

List of Tables

CHAPTER 1: INTRODUCTION

1.1 Background

Heath professional academic programs are under pressure to select, accept, retain, and graduate/train the most qualified students for their respective professions. Applicant numbers however often outweigh the number of available admission spots making the high stakes selection process extremely competitive for those seeking admission into their chosen career path. These applicant numbers in turn pose a daunting and time consuming task for academic programs to not only implement selection procedures, but more importantly utilize reliable and validated assessment measures that allow them to differentiate among candidates, selecting the best for admission.

A psychometric framework is therefore important to ensure that data is sufficiently trustworthy to guide our thinking, and allow appropriate interpretations and accurate decisions (Norman & Eva, 2010) in selection procedures. Determining the various aspects of utility including reliability, validity, feasibility and acceptability to assess the quality of assessment processes is therefore essential.

According to Streiner, Norman & Cairney (2015), reliability is a way to reflect the amount of error (random and systematic) in any measurement. It is also the ability to consistently discriminate how much of the variability observed in scores can be attributed to the actual differences between candidates. Understanding the validity, or whether the selection tool or procedure used is actually measuring what is intended (Streiner, et al., 2015) is also important.

Implementing procedures that are feasible, or practical to undertake with sufficient resources to support using them, as well as acceptable among stakeholder groups to implement, are also key factors to be considered.

Grade point average for example has been shown to be the best predictor of academic achievement in the health professions (Salvatori, 2001), and through meta-analysis has been confirmed to consistently show statistically significant, positive predictive correlations with future performance (Kreiter & Kreiter, 2007; Siu & Reiter, 2009). However, a recent article by Patterson and colleagues raises concern that the discriminatory power of prior academic attainment may be diminishing as the number of applicants with top grades is increasing (Patterson, Knight, Dowell, Nicholson, Cousans, Cleland, 2016).

There is "widespread agreement that it is desirable to broaden the scope of assessment beyond academic achievement" (Eva, et al., 2009, p. 768) as it is necessary for health professionals to not only have the necessary knowledge and cognitive abilities to perform their job, but to also have valuable personal/professional characteristics (e.g. ethical decision-making, relate to others on a personal level, communication skills). This balance of abilities will ensure their ability to collaborate with and relate to patients, colleagues and team members in the delivery of quality patient care. Those candidates with valuable personal/professional characteristics could be excluded if the program's mission is only to base selection on measures of knowledge. Determining which qualities to measure can however be challenging when the profession of medicine has for example

identified 87 positive qualities that are important to superior physician performance

(Price, Lewis, Loughmiller, Nelson, Murray, & Taylor, 1971).

Prideaux and colleagues (2010) recommend that the principles of good assessment

be adopted

> "in defining the purpose of selection; blueprinting of assessable domains and attributes....and utilizing a multi-method programmatic approach in collecting, analysing, interpreting and reporting data from a range of selection instruments, which are fit for purpose" (p.221)

As Siu and Reiter (2009) describe, a medical school would not after all admit students

based on one course's GPA.   A number of screening procedures are commonly used

across health professions in an attempt to assess applicant's personal/professional

characteristics. Standard non-academic requirements typically include letters of reference,

personal statements, and personal interviews (Jones, Simpkins, & Hocking, 2013; Kreiter

& Axelson, 2014). Despite studies noting their inadequate scientific rigor (Jones et al.

2014), a comparison of North American medical schools admissions practices and those

used 25 years ago suggest that little change has taken place in their use during this quarter

century (Kreiter & Axelson, 2013). Few other methods are available however, or

achievable with limited resources (human resources and budgetary) common with many

smaller programs. Preferences of key stakeholders, political considerations, and the

conservative approach often taken by admissions programs can also lead to maintaining

the status quo (Kreiter & Axelson, 2013).

Letters of Reference

Letters of reference are historically one of the most common methods of assessing an applicant's suitability (DeLisa, Jain & Campagnolo, 1994). Letters of reference can however be difficult to interpret when evaluating the quality of one applicant's letter against those of another applicant (Albanese, Snow, Skochelak, Huggett & Farrell, 2003), particularly when letters are more free form in nature with limited structure. The ETS Personal Potential Index has attempted to standardize reference letters for graduate and professional schools, by providing a web-based applicant-specific evaluation of core personal attributes such as communication skills, knowledge and creativity, and ethics and integrity for example for evaluators (ETS Personal Potential Index, 2016). Multiple evaluators can be used per applicant to rate and comment on attributes identified as critical for success in graduate school. ETS found that ratings still suffered "exceedingly from a ceiling effect" (Kyllonen, n.d., p.7). After deleting several items and modifying the wording of several, it is noted that they "expect to demonstrate that the items and scales are sufficiently reliable to support continued use, and...expect psychometric evidence or the validity of the system" (Kyllonen, n.d., p.7-8). Their website however indicates that the ETS Personal Potential Index will be discontinued effective July 1, 2016 without explanation.(ETS Personal Index, 2016).

With standard written letters of reference, an admissions team member's previous experience with an individual reference letter writer can create an additional dimension of challenge when interpretations may be made regarding what is said, or perhaps not said about an applicant. Given that applicants self-select their references, not choosing them

for "their ability to remain neutral and objective" (Siu & Reiter, 2009, 763), their authors tend to give glowing reports, making letters less discriminatory and true opinions about the applicant less clear (Dirschl & Adams, 2000). Add to this challenges with inter-rater reliability (Dirschl & Adams, 2000), poor predictive validity (Kirchner & Holm, 1997; Standridge, Boggs & Mugan, 1997) and rater bias due to open file reviews, (Siu & Reiter, 2009) and there is little evidence to support their effectiveness.

Personal Statements

There is also little psychometric support for the predictive validity of personal statements (Siu & Reiter, 2009). Low inter-rater reliability (Kulatunga-Moruzi, & Norman, 2002), rater bias, and the difficult and subjective comparison of these statements across applicants suggest they add little value to an application (Kulatunga-Moruzi & Norman, 2002). When one additionally factors in the common tendency for these letters to be reviewed and edited by others prior to submission, "questions about help received in its preparation limit the confidence that admission committees can place in its accuracy for all applicants" (Albanese, et al., 2003, p. 318).

Personal Interviews

Personal interviews are also utilized widely as part of the admissions process for health profession education (Goho & Blackman, 2006) for assessment of non-cognitive variables such as professionalism, and communication skills, of the applicant (Goho & Blackman, 2006; Joyner, Cox, White-Harris & Blalock, 2007). Traditional interviews, normally consisting of a panel of interviewers, have demonstrated inadequate reliability

and questionable validity for high stake selection processes for a variety of reasons. Limitations have included: variability in interviewer skill, biases and/or stereotyping by the interviewer (applicant's race, appearance, sex, contrast to other interviewees, or other factors irrelevant to the applicant's suitability), and interviewer scoring tendencies such as central tendency bias (evaluating most candidates as average and not utilizing the range of points on either end of the rating scale), or exhibiting leniency or severity when scoring candidates, (Uijtdehaage, Doyle, & Parker, 2011, Fraga, Oluwasanjo, Wasser, Donato, & Alweis, 2013; Harasym, Woloschuk, Mandin, & Brundin-Mather, 1996; Edwards, Johnson, & Molidor, 1990; Elam & Andrykowski, 1991; Kreiter, Yin, Solow, & Brennan, 2004; Jones, & Forister, 2011).

Despite this finding, programs continue to use traditional interviews. The use of traditional interviews has been so longstanding it can be hard to change this customary and established process. Programs have noted maintaining traditional interviews to acquire more information to clarify the written application, to predict applicant success (Johnson & Edwards, 1991) and assist in the attempt to select candidates who will become good clinicians (Roberts, Walton, Rothnie, Crossley, Lyon, Kumar & Tiller, 2008). However, the exploratory meta-analysis by Goho and Blackman (2006) suggests that selection interviews have only a modest capacity in predicting clinical performance, and only weakly predict academic performance. This might in part be because a personal interview is prone to the bias of context specificity (Eva, et al., 2004a). In an interview, the context (e.g. the interviewer asking the questions and specific questions asked) can play a significant role within which a candidate's performance and responses are assessed

(Eva & Reiter, 2004b). Interviews tend to focus on specific aspects and generalize the appropriateness of that response to all contexts the applicant may encounter. An individual's performance in one situation, or on a particular problem, is however only weakly predictive of the same individual's performance on a different situation or problem (Eva, 2003). Interview results may be further confounded with halo effect, where performance on first questions influence rater's perceptions of candidate performance on subsequent questions. The initial overall impression of a candidate can therefore determine the scores given to individual questions rather than the individual questions adding up to provide a global assessment (Dore, Hanson, Reiter, Blanchard, Deeth, Eva, 2006).

In such a high stakes process where programs are screening in and screening out academically qualified candidates who are eager to train for their chosen career path, it is only reasonable to assume that performance anxiety and nerves may influence a candidate's interview performance. Add to this the potential impact of sequencing effect where a candidate may look better or worse depending on the random occurrence of several good or low quality candidates preceding them than had the candidate had a mixed group before them (Edwards, et al., 1990); the bias that can occur when interviewers review pre-interview information (Gabard, et al., 1997); the finding that excellence or fault in one area can be generalized to imply success or failure in other unrelated areas (Gabard, Porzio, Oxford & Braun, 1997); and unfavourable information can be given more weight than favourable information (Edwards, et al., 1990); and it leads one to further question the utility of a single interview context. Such findings

highlight that a single interview with multiple questions may not provide programs with an accurate, generalisable representation of a candidate's true abilities (Eva, et al., 2004a; Humphrey, Simon, Wall, Diwakar & Goodyear, 2008). The evidence of insufficiency to establish reliability of the traditional preadmission interview (Kreiter, et al., 2004; Kreiter & Axelson, 2013), and its ineffectiveness in predicting important outcomes (Harasym, et al., 1996; Kulatunga-Moruzi & Norman, 2002) call its continued use further into question. As Kreiter and Axelson (2013) write;

> "Clearly, the traditional interview should not be an influential component in selection, and the use of an interview score to make the final decision may violate an applicant's expectation of fair and valid assessment practice" (S52).

The Multiple Mini Interview

To provide a psychometrically fairer assessment of potential performance in admissions contexts, Eva and colleagues (Eva, et al., 2004a) suggest that results should be averaged over many observations of performance to overcome measurement error, just as scores are routinely averaged when conducting knowledge tests (Eva, et al., 2009). The introduction of multiple mini-interviews (MMIs) by researchers Eva, Rosenfeld, Reiter and Norman (2004a) at McMaster University enable a multiple independent sampling methodology (Knorr & Hissbach, 2014) whereby multiple observations of candidates are collected through a series of short independent interviews, usually lasting 8-10 minutes each, to discuss a scenario or respond to questions (Lemay, et al., 2007). This occurs in separate rooms (stations), with a different interviewer in each, to rate candidates instead of having one single panel interview to assess candidates (Eva & Macala, 2014). The MMI format is therefore both defined and strengthened by repeated sampling of

performance (Eva, Reiter, Rosenfeld, Trinh, Wood, & Norman, 2012) across different contexts to dilute the effect of chance and interviewer/situational biases (Eva, et al., 2004a) previously discussed.

This averaging of candidate results across many observations of performance is critical, as this has been found to overcome measurement error (Eva, et al., 2009) and can "[allow] very good candidates to overcome a single poor interview and [prevent] weaker candidates from gaining entry based on a single superior interview" (Eva, et al. 2004b, p.167). The reliability of any one station is consistently low, typically <0.25, while the reliability of the average performance across 12 stations has a median reliability of 0.73 (Eva et al, 2009). This larger number of independent observations provides a more accurate picture of a candidates' strengths and weaknesses (Eva et al, 2004, Eva & Reiter, 2004b) and has been shown to more accurately predict future performance (Eva, Reiter, Rosenfeld, Norman, 2004c; Reiter, Eva, Rosenfeld, Norman, 2007; Eva, et al., 2009; Eva, et al., 2012; Husbands & Dowell, 2013;Hofmeister, Lockyer, Crutcher, 2009; Meredith, Dunlap, Baker, 1982).

With the growing amount of literature supporting the reliability of the multiple-mini interview, this format has been implemented in some medical schools (Eva, et al., 2004a; Eva, et al., 2009; Harris & Owen, 2007; Roberts, et al., 2008) and residency programs (Dore, Kreuger, Ladhani, Rolfson, Kurtz, Kulasegaram, Cullimore, Norman, Eva, Bates, Reiter, 2010; Hofmeister, et al. 2008;) as well as some in rehabilitation sciences (Oxford Grice, 2014), nursing (Perkins, Burton, Dray, Elcock, 2013), midwifery (Callwood, Cooke, Allan, 2014), physician assistant (Jones & Forister, 2011) and

dentistry programs (Roberts, Zoanetti, Rothnie, 2009), and veterinary medicine (Hecker & Violato, 2011).

Uptake of MMI

While the improved psychometric properties of the MMI compared to the traditional interview has led to the use of this candidate assessment tool in the varied professions noted above, it has had a low uptake (Hanson, Kulasegaram, Woods, Fechtig, & Anderson, 2012) across institutions and academic programs. In a 2008 survey of North American medical schools using the Medical College Admission Test (MCAT) (to which 90% of admissions officers responded), Dunleavy and Whittaker (2011) report 83% continue to conduct one-on-one interviews. This is a finding similar to 20 years ago (Edwards, et al., 1990), although results indicate these admissions interviews are more structured than in the past.

Contributing to the low uptake of this multiple independent sampling method is the high resourcing requirements needed to develop and implement 10-12 stations (e.g. development costs, human resources, logistics of large numbers of applicants, interviewer/raters and station rooms needed in one location simultaneously), in addition to the potential effects on recruitment for those schools that use the pre-admission interview and campus visit to familiarize applicants to the positive attributes of the school and faculty (Axelson & Kreiter, 2009). While these can be concerns for large programs, those small in size can find the resourcing burden much greater to adopt the MMI for selections despite their vested interest in using reliable methods to select the most qualified to their programs.

Child Life Studies Program Applicant Assessment Methods

The Child Life Studies Program at McMaster University is one such small program. It is currently one of two child life specialist training programs in Canada. Resources available within this cost-recovery program are small, with respect to budget, faculty, and space. Student tuition and application fees are the only annual source of revenue to support operating costs, and the positions of three part-time faculty and any additional sessional instructors/tutors that may be needed.

Despite these limited resources, it remains the program's mandate and obligation to assure an equitable admissions process, informed by the best evidence. As a small, but growing profession, there has been a steady increase in the number of applicants applying for admission, increasing from 55 in 2005 to 141 in 2015. The number of candidates offered admission has however remained stable at 10 each year.

Differentiating among candidates to select the best is critical within this profession. Strong personal and professional characteristics are needed for child life interns and graduates to engage with children, youth and families amid pediatric health crises, and stressful life experiences (e.g. new chronic or life-limiting diagnoses, traumatic injury, death of sibling). While the ability to assess potential and current risks on a child and family's resiliency to develop individualized interventions is important, the ability to effectively implement these interventions through supportive and therapeutic relationships is key to promote and support children's coping and adjustment.

Pre-Interview Selection

With few options available to assess personal/professional characteristics to differentiate among applicants, modifications, informed by evidence, were undertaken in an attempt to structure a better selection process and enhance the psychometric utility. Two application requirements do continue to be received without any modification to screening procedures. Applicant CVs continue to be reviewed to capture applicant experience, and GPA remains an important screening tool which aligns with the literature previously discussed.

Given the resource limited nature of the Child Life Studies program, modifications to maximize the utility of the selection process have been made within resource capacity. The first such modification included a change in the requirements of the pre-interview written submission/personal statement. Applicants are asked to address three questions in their statement and submit as part of the original application submission. This resulted in more structured letters to evaluate rather than when they were free form in presentation and consequently more difficult to compare across applicants. A horizontal scoring and review process has also been initiated whereby the three questions within the written submission are rated independently (e.g. raters review all question ones before moving to question twos), so performance on previous questions do not create a halo effect and influence a rater's impression on subsequent questions (Dore, et al., 2006). With this method, more observations can be attained to provide a global assessment (Dore et al., 2006). Once the applicant letters have all been evaluated, the remainder of the application is reviewed and scored.

The information quality within reference letters can vary substantially depending on the degree of structure provided to the letter writers (Kreiter and Axelson, 2013). As a second modification to applicant screening procedures for the Child Life Studies program, a more structured format has been put in place for the two required letters of reference. Referees are asked to use narrative descriptions to address specific questions provided within a referee form, and also use ratings to address a set of personal/professional characteristics such as communication skills, ability to work independently and with others, dependability, empathy, and professional attitude, among others. These characteristics were chosen, as rating high in these areas is deemed important for child life clinical practice. This additional structure serves to provide referees with a more standardized method of providing performance evaluation. This specificity also provides the program with useful comparators across applicants vs. when reference letters were free form in nature.

While some programs have chosen to discontinue their personal interviews due to the time commitment required of faculty and expense to the university (Gabard, 1997), the McMaster Child Life Studies program has continued traditional panel interviews as a means of assessing personal/professional skills. Applicant files are selected for interview based on GPA, personal statement, reference letters and relevant experience with children, youth and families. Once a decision to interview a candidate is made, the applicant file is not reviewed again as the availability of pre-interview information on candidates could have bias effects (Goho & Blackman, 2006).

Panel Interview Process

From 1989 until 2013, the interview(s) themselves were as structured as possible to reduce bias (Pau, Jeevaratnam, Chen, Fall, Khoo, & Nadarajah, 2013). Panel interviews were booked for 1 hour blocks with an additional 15 minutes after completion of the interview for raters to score candidates (see Appendix A for full panel interview schedule). A total of 24 candidates were interviewed with six interviews held each day, across the four days. This method therefore required five hours of interviewer time per candidate (1 hr. and 15 mins. interview x 4 interviewers) for a total of 120 hours of interviewer time over four days.

Interviews were all conducted by the same three faculty members who asked all applicants the same questions. Additionally, a recent graduate was part of each interview (one for each of the four days), who received training in advance on the interview process. Standard procedure included two interviewers on the panel remaining in the room with the candidate throughout the interview, and two sitting behind a two-way mirror. The recent graduate involved each day also had the opportunity to interview candidates in the room after several observations through the two-way mirror.

Candidates completed their one-hour scheduled interview at varying times. Following the individual interview, the full panel reassembled so candidates were able to ask any questions they had about the program in the time remaining. Each interviewer scored candidates independently. Group consensus was not required for candidate selection as final interview scores were averaged. While there was evidence to support strong inter-rater reliability, Razack, and colleagues (2009) note that "although inter-rater

reliability with multiple interviewers in the same interview can be good, it tends to be poor across interviews" (p.994).

Despite attempts to enhance structure and reliability across application screening measures, this program observed and anecdotally reported that, in some cases, professionalism, the ability to receive and accept constructive feedback and collaborate within a team of peers were not always discernible. Several students gained admittance with less than desirable levels of these personal/professional characteristics, resulting in high resource usage from the program in an attempt to provide remediation. After the program's twenty-three year history of conducting panel interviews for admission, maintaining this precedent was not justifiable given both the program's observations and the literature demonstrating that panel interviews produce a score with low to moderate reliability, resulting in zero to low validity (Kreiter, et al., 2004, p. 158).

Census of other small academic programs in child life

Informal outreach and consultation with five other Child Life program directors identified that graduate programs in this field of study (there are only seven graduate degree programs in North America with a child life major) are all similar in size (10-15 students) with very small numbers of faculty, and in some cases greater numbers of sessional or adjunct instructors to deliver their programs. Personal interviews were also used as part of their admissions decisions with several of those consulted indicating they had initiated rolling admissions, interviewing candidates upon application. Two of these graduate programs had a single interviewer conducting one on one interviews with applicants by phone to base admissions decisions. During an on-site program review with

one of these programs in the United States, the Director/candidate interviewer felt she may have better luck selecting candidates if she just threw all of the applications in the air and selected those that landed on the right side of the room (personal communication, 2014). Uncertainty resounded among those consulted about whether they were using the best approach in candidate selection as they experienced more challenges with students and their performance despite minimum GPA entrance requirements.

Candidate selection was therefore assessed to be of importance to academic programs across the child life profession with relevance to other psychosocial based professions and small academic programs who continue to conduct panel or single interviews despite limitations on their efficacy.

Implementation of MMI in McMaster's Child Life Studies Program

While program size and resources continue to remain a factor, the McMaster Child Life Studies program identified that repeated sampling of interviewed candidate's performance across different contexts was needed. In addition to the application screening modifications noted above, a multiple independent sampling methodology was planned through the initiation of multiple-mini interviews. Instead of candidates having one single panel interview, several independent encounters were planned in separate rooms (stations) with one interviewer in each.

The usual number of MMI stations reported in the literature for candidate selection is between six to twelve (Knorr & Hissbach, 2014). Resources within the Child Life Studies program however permitted the development and implementation of a 4-

station MMI in 2014, and 2015. Only three MMI studies have been located with a 4-station design (Finlayson & Townson, 2011; Serres & Peeters, 2012; Hanson, Kulasegaram, Woods, Fechtig, Anderson, 2012). All three studies involved medical students. Two were selecting for residency training programs (Finlayson & Townson, 2011; Serres & Peters, 2012), and the third for a PhD program in Leadership Education and Development (Hanson, et al., 2012) with candidates from their university's medical school. Other aspects of non-academic performance had already been assessed during their university's medical school admissions process (Hanson, et al., 2012). These studies therefore report on the selection procedure for specialty programs. No 4-station design could be found for comparison in the selection of candidates for a health professional academic program.

While the increase in stations were hypothesized to enhance reliability from the current single interview format, how well it actually works and any pitfalls to avoid in the MMI design is the focus of this thesis. This is the first MMI tool to be developed within child life education, and is also thought to be the first in the psychosocial based professions. This study will serve to provide quality assurance and quality improvement in McMaster child life student selections, as well as address a gap in the literature to generate findings for other psychosocial based professions and small programs to consider in their admissions processes.

Study Purpose

This thesis sought to understand the following study questions;

1. Is a 4-station MMI feasible for small programs to undertake?

2. Is a 4-station MMI acceptable among child life candidates and interviewers?

3. Is this 4-station MMI design reliable?

4. Are there any predictors of candidate's MMI performance?

5. Are there pitfalls in small MMIs that can be minimized or avoided?

CHAPTER 2: METHODS

2.1 Phases of the Study

The methods section of this thesis will be broken down into the four phases undertaken in this study. Phase one included development of a 4-station MMI in 2014 and in 2015. Phase two involved implementation of the MMI and implementing stakeholder surveys in 2015 to explore participants' perceptions of MMI, acceptability of its use, and any suggestions for improvement. The third phase involved a review of MMI candidates (n=75) application packages to identify predictors of performance on the MMI. The fourth phase involved statistical analysis of these predictors to determine if there was any relationship to candidate's MMI performance, and the reliability of the 4-station MMI design in 2014 and 2015. It also sought to determine the preferred length of MMI to achieve minimally acceptable overall test reliability, and its feasibility in a small program.

2.2 Sample size and Selection

In 2014, the first year MMI was introduced, a total of 125 applicants were screened through admission procedures. This was a heterogeneous applicant pool with candidates coming from varied educational backgrounds (e.g. Psychology, Child and Youth Studies, Health Sciences, Biology, Education), ages, volunteer and work experience, among others. Some applied from their undergraduate studies, while others applied for a second career as a mature student. A total of 35 were interviewed with the new MMI system reflecting 28% (35/125) of the applicant pool, compared to a total of 24 interviewed the previous year by panel interview reflecting only 22.4% of applicants (24/107). Of those interviewed, 33 were female and 2 were male (5.7%). These 2 males were however the only male applicants to apply (1.6% of the applicant pool).

During the selection of the 2015 cohort of child life students, a similar diverse group of applicants applied. A total of 141 applicants were screened and 40 offered interviews reflecting 28.3% of the applicant pool. Within this group, 39 of those interviewed were female and 1 was male (2.5%).

While this gender imbalance may seem atypical for many health professions programs, this is a common trend within the child life profession. According to the Manager of Professional Resources and Services of the Child Life Council, the child life specialist North American professional association and certification granting body, current member rates show 98.4% of the profession is female (personal communication, 2015). This disparity in gender is therefore commonly seen within the population of applicants, students, and practicing child life professionals.

Table 1 Study Sample Size & Gender Distribution

Year	# of Applicants	Applicant gender distribution	# Interviewed	Applicant gender distribution
2013	107	F= 105 M= 2	24	F= 23 M= 1
2014	125	F= 123 M= 2	35	F= 33 M= 2
2015	141	F= 138 M= 3	40	F= 39 M= 1

All McMaster University Child Life Studies faculty members (n=4) participated as MMI station evaluators. While this sample of interviewers is quite small, it reflects the only Canadian faculty members interviewing child life candidates.

2.3 Ethics

Because this study is deemed as a Quality Assurance Project, REB approval was not needed. All participants were however required to sign consent (see Appendix D and E) prior to completing the stakeholder surveys. All applicant files, MMI score sheets, and stakeholder surveys were kept confidential and locked in the investigator's office.

2. 4 Phase One - Development of a 4-Station MMI

Reiter and Eva (2005) recommend that programs define and assess the relative value of the personal/professional characteristics that society, their institution and the health profession to which the candidate is applying find desirable to guide and develop admissions protocols. These characteristics might include ethical decision making, collaboration and communication among others (Eva, et al., 2004a, Reiter & Eva, 2005).

This information can then assist programs in developing a blueprint of those characteristics to ensure they are representatively sampled through questions or scenarios in the admissions interview process. The use of separate interview stations within the MMI format provides programs with flexibility in ways to develop station formats to assess those characteristics (e.g. scenarios, interacting with a standardized patient or trained actor, traditional questions, and so on (Knorr & Hirshbach, 2014). This program chose to create stations and questions/scenarios addressing the domains of critical thinking, communication, collaboration and professionalism.

Building off of the findings of Eva et al., (2004b), who report, "greater gains are gained in the reliability of a personal interview by increasing the number of stations and decreasing the number of interviewers per station…." (p.167), the four station interview process was designed. The same three faculty members and the program's Clinical Internship Assistant acted as interviewers, each rating one of the four interview stations in separate rooms. Raters remained in the same station throughout the interviews. This was done to adjust for interviewer stringency/leniency (Roberts, Rothnie, Zoanetti, & Crossley, 2010), and increase the likelihood that station scoring would be equitable with comparators set early in the interview process, and to limit the time needed to train interviewers.

Station 1-3

Each station was created to be fifteen minutes in length with an additional five minutes for interviewers to score, and allow applicants to have a break. Stations 1-3 each contained three questions/scenarios, for a total of 9 questions across the 3 stations. In

contrast to larger health professions, such as medicine and nursing, there can still be varying degrees of role ambiguity surrounding the role of a child life specialist by the public, and at times by applicants seeking admission to the Child Life Studies program. Similar to the action research team members in Dotson Collin's 2013 dissertation, child life faculty also believed an awareness and understanding of "what they're signing up for" was important for program attrition rates, for student success, and for patient and family interaction and care. Unlike the other stations, Station 1 included questions and/or situations centred around child life to better understand an applicant's understanding of the role of a child life specialist in healthcare.

A 4-point scale was used by faculty to independently mark performance on each question in her station. The following describes the ratings;

- 1=unsatisfactory

- 2=below average

- 3=good

- 4=outstanding

A 4-point scale has been used with good results in other studies (Roberts, et. al., 2008; Kumar, Roberts, Rothnie, du Fresne & Walton, 2009; Roberts et al., 2009; Roberts, et al. 2010; Jerant, Griffin, Rainwater, Henderson, Sousa, Bertakis, Fenton, & Franks, 2012; Tiller, O'Mara, Rothnie, Dunn, Lee, Roberts, 2013) and was intended to encourage interviewers to choose a specific score, as no indifferent option was available, to "reduce the amount of statistical noise, improve reliability…and allow for better admissions decisions" (Stansfield & Kreiter, 2007, pg. 37). A comment section was also available on

each scoring sheet, providing space to highlight any summative comments for each candidate, or red flag any concerns.

Station 4

Station 4 included a writing sample where candidates were asked to summarize a research abstract in his/her own words. Ten minutes was allotted and timed for this task. Unlike other questions, this was rated out of 20 points. This question score was later averaged to a score out of 4 to align with all other question scores.

Qualitative responses to a survey of interviewed graduate-entry medical school applicants by Kumar and colleagues (2009) identified that candidates felt there was limited opportunity to discuss their specific personal qualities to present a holistic picture of themselves. One additional question was asked in this station for candidate to address in the remaining five minutes. The question "If we have you and one applicant left and only one spot, why should we choose you?" was therefore asked within this station for candidates to self-present values and speak about experiences.

2.5 Phase Two

2.5a - Part A 2014 MMI Interview Schedule and Implementation

In 2014, 36 candidates were invited to interview, twelve per day, across three days. One candidate failed to show for her scheduled interview. A total of 35 candidates were therefore interviewed. During this first implementation of the MMIs, a staggered start approach was utilized (see Appendix B for the 2014 MMI interview station schedule). Each candidate was assigned an individual start time and signed in at a central

check-in desk upon arrival. He/she rotated through the same stations in sequential order. The interview stations and questions remained the same for the three sequential interview days.

All interviews were conducted in person, with the exception of two Skype interviews in 2014. One candidate in Australia and one in France were unable to attend their four station interviews in person. The Child Life Studies Program has experienced globalization of admissions selection with applicants in the program's history from across Canada, the US, Australia, Cayman Islands, Japan, Netherlands, Singapore, Uganda and Qatar. A fifth iPad was therefore made available during the MMI interviews to "…ensure that local and international students are selected using the same criteria providing equity of access" (Tiller, et al., 2013, p. 809). The results of Tiller and colleague's 2013 study show Skype as a favourable option at each interview station in those unique circumstances. They report, "the cost and time savings for candidates were substantial and in [their] view outweighed the desire to conduct interviews in person as demonstrated by the 74% overall satisfaction rate with being interviewed online" (p.809). In these two circumstances, the iPad travelled room to room so the candidates could complete their stations with the interviewers via Skype, and interviewers could maintain independence of scoring.

In the written notification candidates received with his/her interview offer, an invitation was also included to participate in an informal program information and question period (one post-morning interviews, or one post-afternoon interviews). This invitation was extended again personally upon check-in. Both question and answer (Q

&A) sessions were scheduled for half an hour in length to address common questions received by candidates, and discuss the program's academic and clinical components. As Phillips and Garmel (2014) note, there are some who find the MMI to be a unidirectional experience of moving to interview stations without a sense of the interviewers or the program's personality. This Q & A was therefore incorporated for all candidates to address this identified gap, and also maintain the opportunity to address candidate questions as historically done with the panel interview method.

2.5b Part B 2015 MMI Interview Schedule and Implementation

In 2015, 40 candidates were invited to interview, twenty per day, across two days. Instead of a staggered start and working through the stations in sequential order, the candidates from this year were assigned into groups of four. Five groups of four were assigned for both interview days. This increased the number of candidates interviewed in a shorter period of time. Each group was provided a start time, and randomly assigned to a starting station (see Appendix C for the 2015 MMI interview schedule). When the time came for each group to begin the circuit, a volunteer (new graduate) escorted the group to their assigned station (A, B, C or D).

As in 2014, interviewed candidates were invited through the same methods to participate in an informal program information and question period (one post-morning interviews, or one post-afternoon interviews). The Q &A sessions provided the same opportunity for questions and discussion as in the previous year.

2.5c Stakeholder Survey

In 2015, immediately following the four interview stations, and before the outcome of the interviews were known, the interview volunteers asked each group of 4 candidates to complete an anonymous survey in a separate room. Candidates were also invited to participate in the exit surveys in their interview offer letter. Voluntary consent was gained with applicants informed that participation would not have any bearing on his/her admission status, and that no detriment would arise by participating (see Appendix D for the candidate consent form used). The paper-based survey was comprised of questions with a Likert response scale (1=strongly disagree, 7=strongly agree) related to how well organized they found the interviews to be, if they could accurately portray themselves, if candidates found the MMI to be a fair process, among others. The middle position of the scale was labelled 'neither agree nor disagree' to enable a neutral position for candidates, and provide the choice of expressing no opinion. Including this neutral position created seven categories within the scale. Scales with seven categories have been found to be more reliable than those with fewer categories, with evidence to suggest that people are often unable to discriminate beyond seven levels (Streiner, et al., 2015). An open text space was also available for candidates to provide comments and descriptive feedback if they found any limitations to the MMI (see Appendix E for the applicant survey). Completed surveys were returned to the volunteers.

One final paper-based survey was also distributed upon completion of the morning and afternoon program information and question period to evaluate the inclusion of this component in the interview process (see Appendix F for the post program question

and answer period survey). This survey included questions like, "My questions were addressed in this format", and "I have greater knowledge of the Child Life Studies Program and faculty". The same Likert response scale was provided (1=strongly disagree, 7= strongly agree) in addition to an open text space to identify any limitations to the question and answer period. Completed surveys were returned to the graduate volunteers.

Faculty interviewers were also asked to complete an anonymous survey at the conclusion of the two interview days with questions related to things like the effectiveness of the MMI process and its feasibility (see Appendix G for the interviewer consent form and Appendix H for the interviewer survey). The same Likert based scale was utilized with completed surveys also returned to the volunteers.

2.6 Phase Three - Predictors of MMI Performance

A review of all interviewed candidates' applications from 2014 (n=35) and 2015 (n=40) was undertaken to identify candidates' age, professional and volunteer work experiences in addition to whether they had previous child life course work and experience with children with disabilities. A t-test was performed for each of these variables to determine whether there were significant differences in MMI score between candidates with and without these variables to determine if there were any predictors in candidate's MMI performance. Candidate degree type was also identified for all candidates interviewed and a one-way ANOVA performed to see if there were any differences in MMI performance based on candidate's degree type. A descriptive analysis of this demographic information was also used to discern the relative homogeneity or heterogeneity of the pool of interviewed candidates for both 2014 and 2015.

2.7. Phase Four - Statistical Analysis

Psychometric methods have been used to evaluate whether the data gathered is trustworthy enough to allow appropriate interpretations and accurate decision making on the use of a 4-station MMI (Norman & Eva, 2010). An evaluation of feasibility, acceptability and reliability has been undertaken to determine the usefulness of the MMI as an assessment method.

A review of the feasibility of developing and implementing a 4-station MMI is considered by comparing time and costs to the panel interviews previously used to determine cost efficiency. The more practical it is to use within applicant selections, the more feasible it will be to use.

As noted in phase 3 of this study, the level of acceptability among faculty and candidates on the usefulness of the MMI for applicant selections was collected through stakeholder surveys. Their levels of acceptability were assessed by analyzing the scores of satisfaction on each form completed, and in reviewing respondents' descriptive feedback.

In order to determine if a 4-station MMI is reliable, we must determine if it is able to differentiate among candidates, or more specifically consistently differentiate strong candidates from weaker ones. Generalizability coefficients were calculated to determine if the variability in scores reflect true differences between candidates assessed through the MMI. By understanding the interaction effect between candidates and how they differ over stations and questions in both 2014 and 2015, we can determine how reliable the scores obtained within the MMI are.

Another important aspect of psychometric analyses is to determine how many observations one must collect in total to achieve reasonable levels of reliability (Norman & Eva, 2010). Analyses for the 2014 and 2015 data sets therefore include predicted levels of reliability if more observations were collected through adding more MMI stations and questions. Correlations to assess the relationship between overall station scores were also calculated using the Statistics Package for Social Sciences (SPSS) software.

Certification scores could not be provided by the North American child life professional association. MMI could therefore not be used as a predictor variable and post-school certification scores as outcomes variables to evaluate predictive validity of a 4-station MMI within this thesis.

CHAPTER 3: RESULTS

3.1 Feasibility

Part A. 2014 Feasibility Results

The time and costs associated with implementing the MMI on the day of the interview are compared to the previously used panel interview format. Table 2 provides a summary of the staff time spent in the panel interview format (120hrs.) compared to time spent in the MMI implementation and program question and answer period in 2014 (72.75 hrs.). There was an increase seen in the number of candidates interviewed within a shorter time period (35 in 3 days in 2014, compared to 24 over 4 days using the panel format). More volunteer support was however required to facilitate the MMI interviews, captured in Table 2.

A greater number of situation-based questions were used in 2014 requiring more staff time to develop. More staff time was also required to coordinate multiple station/interview spaces in close proximity, although this time was not formally tracked. Less staff time was however needed to address candidate inquiries following the incorporation of the Question and Answer period than when panel interviews were offered. While the specific number of inquiries received and hours spent addressing them in previous years were not tracked, more staff time was required to address candidate inquiries via email or phone following the panel interviews compared to the MMIs.

Table 2 Comparison of Time Spent in previous panel interviews and the 2014 MMI design.

Interview type	Number of candidates Interviewed	Staff time per candidate	Time needed for group Q&A sessions	Calculation of total staff time to interview
Traditional Panel Interviews	N= 24 6 candidates x 4 days	1hr. interview by panel of 4 + 15 mins. to score = 5 hrs. of *staff time* per candidate **Total 120 hrs**.	No formal group Q&A. Questions answered if time left. Many questions received via email post-interview.	**120 hrs of** *staff* **time over 4 days** + Additional time needed to respond to email inquiries.
4 Station MMI Candidates have staggered start and move through stations in progressive order	N=36 12 candidates x 3 days	20 minutes per station x 4 stations = 1hr. and 20 mins. of *staff time* per candidate 6.25hrs. x 3 days = 18.75hrs. of *volunteer* time to man check-in desk **Total 66.75 hrs**	30 mins. x 3 days x 4 interviewers **Total 6 hrs**	**54 hrs of** *staff* **time over 3 days** **18.75 hrs. of** *volunteer* **time over 3 days** **Total 72.75 hrs over 3 days**

Part B. 2015 Feasibility Results

Table 3 provides a summary of the staff time spent in the MMIs and program question and answer period in 2015. A total of 40 candidates were interviewed over two days as five groups of four rotated through the interview stations on both days. The same raters maintained their stations throughout, and the additional staff/volunteers oversaw the check-in desk, acted as timekeepers and distributed and collected stakeholder surveys. Like 2014, the time spent developing station questions was not tracked, but less time was used during this year with the incorporation of more behavioural interviewing questions.

Table 3 Staff Time Spent in the 2015 MMI design

Interview type	Number of candidates Interviewed	Staff time per candidate	Time needed for group Q&A sessions	Calculation of total staff time to interview
4 Station MMI. Candidates assigned to 1 of 4 stations in random order	N=40 20 candidates x 2 days	20 mins. per station x 4 stations = 1 hr. and 20 mins. of *staff time* per candidate 2 volunteers/day x 9hrs/day x 2 days = 36hrs. of *volunteer* time to man check in desk, time-keep & facilitate stakeholder surveys **Total 89 hrs**	30 mins. x 2 days x 4 interviewers **Total 4 hrs**	**57 hrs. of *staff* time over 2 days** **36hrs. of *volunteer* time over 2 days** **Total 93 hrs. over 2 days**

3.2 Acceptability

3.2a Candidate post-MMI exit survey

All forty candidates completed the post-MMI exit survey (100% response rate) immediately following the completion of their 4 stations, and returned them to the interview volunteers. Table 3 shows the scaled responses received, and the mean calculations for each survey question. Mean question scores range from 6 to 6.8 on the 7-point Likert scale. A summary of the survey respondent's descriptive feedback follows to illustrate major themes in survey participant's perceptions of MMI, acceptability of its use in child life applicant selection, and in identifying any suggestion for quality improvements.

Table 4 Candidate Post MMI Exit Survey Data n=40/40 (100%) respondents

(1- strongly disagree, 2- disagree, 3 - disagree somewhat, 4- neither agree nor disagree, 5- agree somewhat, 6- agree, or 7- strongly agree)

Question	1	2	3	4	5	6	7	Mean
1. The interview process was well organized.						8(20%)	32(80%)	6.8
2. I had sufficient time to present my ideas in the MMI.					3(7.5%)	20(50%)	17(42.5%)	6.35
3. I could accurately portray myself in the MMI.			1(2.5%)	1(2.5%)	6(15%)	21(52.5%)	11(27.5%)	6
4. I personally prefer meeting 1 on 1 with an interviewer in					1(2.5%)	6(15%)	33(82.5%)	6.73

each station instead of with a traditional interview panel.							
5. Compared to traditional panel interviews, I found the MMI was more fair.		1(2.5%)	2(5%)	4(10%)	15(37.5%)	18(45%)	6.15
6. I enjoyed participating in the MMI.				1(2.5%)	20(50%)	19(47.5%)	6.45
7. I am satisfied with the MMI as one of the tools for the admissions selection process into the Child Life Studies Program.				3(7.5%)	21(52.5%)	16(40%)	6.33

Adapted from Oxford Grice (2014), Humphrey, et al. 2008, Razack, et al., 2009

The question, "Did you find any limitations to the MMI?" followed with an open text response format. Several indicated they would have liked more information ahead of time on the types of questions to be asked to help prepare for the interview. Other comments included wishing there was more of an opportunity to express who they were (e.g. experiences, backgrounds, strengths). Nine of 40 (22.5%) indicated they wished they had more time. This however was tied to the Station 4 writing component.

3.2b Post-Program Q & A Exit Survey Data

All 40 candidates (20 per day) were invited to stay for the program question and answer (Q & A) sessions. The first 3 rounds of candidates (n=12) were invited to join the morning Q & A session both days, and the final 2 rounds of candidates (n=8) were invited to join the afternoon session both days. Thirty-three of forty candidates returned for the Q & A sessions over the 2 interview days, and completed a survey immediately following. Table 4 shows the scaled responses received, and the mean calculations for each survey question. Mean scores range from 6.55 to 6.69 on the 7-point Likert scale. A summary of the survey respondents' descriptive feedback follows to illustrate major themes in whether they found any limitations.

Table 5 Candidate Post Q & A Exit Survey Data n=33/33 (100%) respondents

(1- strongly disagree, 2- disagree, 3 - disagree somewhat, 4- neither agree nor disagree, 5- agree somewhat, 6- agree, or 7- strongly agree)

Question	1	2	3	4	5	6	7	Mean
The program question and answer period was informative.						11(33.33%)	22(66.67%)	6.67
My questions were addressed in this format.						15(45.5%)	18(54.5%)	6.55
I have greater knowledge of the Child Life Studies program and faculty.						10(30.3%)	23(69.7%)	6.69

The question "Were there any limitations to this question and answer period?" included an open text response format. Many comments were received about it being helpful, casual and informative and finding it beneficial to be with others asking questions

to enhance learning. One candidate suggested "having more grad. students, or additional meet & greet, less formal Q & A session with recent graduates". In addition, one candidate commented that "with such a large group, may not have asked questions".

3.2c Interviewer Exit Survey

The four interviewers in our 4-station MMI who remained in their stations for the two interview days completed an exit survey at the conclusion of day two. All four (including this writer) completed the questionnaire. Table 5 shows the scaled responses received, and the mean calculations for each survey question. Mean scores range from 6-7 on the 7-point Likert scale. A summary of their descriptive feedback follows and illustrates major themes in whether they found any limitations or advantages to the MMI.

Table 6 Interviewer Post MMI Exit Survey Data n=4/4 (100%) respondents

(1- strongly disagree, 2- disagree, 3 - disagree somewhat, 4- neither agree nor disagree, 5- agree somewhat, 6- agree, or 7- strongly agree)

Question	1	2	3	4	5	6	7	Mean
Interview stations are of the right duration.					1(25%)	2(50%)	1(25%)	6
We are measuring the right skills and abilities.						1(25%)	3(75%)	6.75
I had time to score applicants before the next applicant came into the room.					1(25%)		3(75%)	6.5
The scoring sheet allowed me to differentiate among applicants.			1(25%)				3(75%)	6
I would consistently offer the same assessment if asked to						2(50%)	2(50%)	6.5

review my opinion at a later time.								
My interviewer workload is less in MMIs than traditional interviews.							4(100%)	7
I personally prefer meeting 1 on 1 with an applicant in each station instead of being on a traditional interview panel.							4(100%)	7
Compared to traditional panel interviews, I found the MMI was more fair.							4(100%)	7
I would be in favour of continuing to utilize the MMI instead of traditional interviews.							4(100%)	7

Adapted from Humphrey et al., 2008, Razack et al, 2009, Dore et al., 2010

The question "Did you find any limitations to the MMI?" included an open text response format. One interviewer found in her station (all interviewers were female) that a few could have used more time, yet another interviewer thought the stations could be shorter as she typically had an extra two to three minutes on average at the end of her15 minute station. She suggested more prompts provided might help to "extend the questions/get more out of the candidates if they are really nervous". One additional comment included, "As with any questions, sometimes applicants got off topic".

The survey also asked interviewers "Did you find any advantages to the MMI?". Comments included feeling like the candidates were less anxious with just 1 interviewer, and this process led to acquiring more pertinent information from candidates. One interviewer noted that it seemed like candidates had a fresh start when coming into her

station. There was a consistent theme that this format enabled seeing more candidates in less time which lessened interviewer fatigue and provided a "huge cost benefit from a human resource perspective and cost".

3.3 Reliability

Reliability was calculated with generalizability (G) theory through urGenova software. One of the advantages of G-theory includes the ability to estimate the degree of various sources of error and their relative importance in a single analysis (Sebok, Luu, Klinger, 2014). A generalizability or G-study was performed resulting in a G coefficient to estimate variation due to candidates, and the contributions of different sources of variance. The G-coefficient therefore provides a measure of how confident one can be that any differences detected between candidates are real differences (Crossley, et.al, 2002). Candidates (c) are included as a source of variance, and station (s) and question nested within station (q:s) are the major sources of error, or facets of interest in the G study.

The data sets for 2014 and 2015 include the scores each candidate received per station question. A total of 11 scores are included (three in station 1, three in station 2, three in station 3 and two in station 4).

As previously discussed, Station four included the writing question which was rated out of 20 points. This was averaged to a score of four to appropriately weight this as one question in the statistical analysis. By doing so, all individual data items are on the

same identical numerical scale. An overall MMI score was calculated out of 16 for each candidate.

3.3a Part A 2014 Reliability Results

The 2014 G-study yielded overall test generalizability equal to 0.718. Table 7 also reports the variance components. A higher variance between stations (0.185) was noted than between candidates (0.089) and questions within stations (0.036).

Table 7 Summary of estimated Variance Components, effects and the G-coefficient (Experiment 1 - 2014)

Facet	Df	MS	Variance
Candidate	34	1.359	0.089
Station	3	19.317	0.185
Question nested in Station	7	1.597	0.036
G-coefficient			0.718

A decision or D-study was also performed to predict reliability across different situations (Crossley, Davies, Humphris, Jolly, 2012) and determine how alterations to the number of observations (stations and question) collected could improve levels of reliability and yield the highest generalizability (Bloch & Norman, 2012). Analyses for the data set included what might be expected of an 8-station (0.847) and 12-station (0.892) MMI. It also included what might be expected if station 4 were modified to include three questions (0.734), instead of the current two, in the 4-station MMI design. Table 8 provides these results for 2014.

Table 8 D Study - 2014 Predicted reliability if increased stations

Number of Stations	Number of Questions per Station	D-Study Adjusted Co-efficient
4	3	0.734
8	3	0.847
12	3	0. 892

The range of scores, mean and standard deviation for the MMI questions and station averages are included in Table 9. Station 1 yields the highest mean scores with the station average at 3.543. The Station 2 average is 3.028. Station 3 and 4's averages are 2.571 and 2.666 respectively.

Table 9 2014 Range of Scores, Mean & Standard Deviation for MMI Questions & Station Averages

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Station 1 Question 1	35	2	2	4	3.540	0.611
Station 1 Question 2	35	2	2	4	3.400	0.553
Station 1 Question 3	35	1	3	4	3.690	0.471
Station 1 Average	35	1.667	2.333	4	3.543	0.428
Station 2 Question 1	35	2	2	4	3.030	0.618
Station 2 Question 2	35	3	1	4	3.060	0.906
Station 2 Question 3	35	2	2	4	3.000	0.642
Station2 Average	35	1.667	2.000	3.667	3.029	0.467
Station 3 Question 1	35	3	1	4	2.170	0.891
Station 3 Question 2	35	2	2	4	2.690	0.676
Station 3 Question 3	35	2	2	4	2.860	0.550
Station3 Average	35	2.333	1.667	4	2.572	0.502
Station 4 Question 1	35	2	1.600	3.600	2.560	0.578
Station 4 Question 2	35	3	1	4	2.770	0.690
Station 4 Average	35	2	1.800	3.800	2.666	0.521

The mean and standard deviation of question scores and station averages were also completed to compare scores between the three different interview days held in 2014. The results are reported below in Table 10.

Table 10 2014 Mean & Standard Deviation for MMI Questions & Station per Interview Day

	Day 1 Mean	Day 1 Std. Deviation	Day 2 Mean	Day 2 Std. Deviation	Day 3 Mean	Day 3 Std. Deviation
Station 1 Question 1	3.55	0.82	3.5	0.52	3.58	0.51
Station 1 Question 2	3.3	0.6	3.3	0.5	3.6	0.5
Station 1 Question 3	3.8	0.4	3.6	0.5	3.7	0.5
Station 1 Average	3.545	0.563	3.472	0.388	3.611	0.343
Station 2 Question 1	3	0.77	3	0.43	3.08	0.67
Station 2 Question 2	3	0.9	3.2	1	3	0.9
Station 2 Question 3	2.91	0.54	2.92	0.67	3.17	0.72
Station 2 Average	2.970	0.586	3.028	0.481	3.083	0.352
Station 3 Question 1	2.45	1.04	2	0.6	2.08	1
Station 3 Question 2	2.64	0.67	2.42	0.67	3	0.6
Station 3 Question 3	2.82	0.6	2.75	0.45	3	0.6
Station 3 Average	2.636	0.623	2.389	0.372	2.694	0.481
Station 4 Question 1	2.273	0.561	2.617	0.581	2.767	0.525
Station 4 Question 2	2.727	0.647	2.667	0.778	2.917	0.669
Station 4 Average	2.5	0.452	2.642	0.598	2.842	0.483

Pearson correlations were also used to assess the relationship between overall station scores. Results of the 2014 data are shown in Table 11. There was a significant correlation between all stations.

Table 11 Pearson Correlations for 2014 MMI Station Average Scores (n=35)

2014		Station 1 Average	Station 2 Average	Station 3 Average	Station 4 Average
Station 1 Average	Pearson Correlation	1	0.459**	0.627**	0.407*
	Sig. (2-tailed)		0.006	0	0.015
Station 2 Average	Pearson Correlation	.459**	1	.360*	0.395*
	Sig. (2-tailed)	0.006		0.033	0.019
Station 3 Average	Pearson Correlation	.627**	.360*	1	0.043
	Sig. (2-tailed)	0	0.033		0.805
Station 4 Average	Pearson Correlation	0.407*	0.395*	0.043	1
	Sig. (2-tailed)	0.015	0.019	0.805	

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

3.4 2014 Predictors of MMI Performance

The applications of the 35 candidates interviewed in 2014 were reviewed to collect their background information and analyze the relative homogeneity or heterogeneity of the pool of interviewed candidates. This information was also used to discern if there were any predictors of candidate performance on the MMI based on their experience in health care, professional experience, experience with children with disabilities, and/or having taken previous child life course work. Each predictor will be reported in the sequence identified above.

3.4a. 2014 Candidate Healthcare Experience

An independent sample t-test was conducted to assess whether the mean scores of candidate's with experience in healthcare (mean= 11.848, SD = 1.456) and those without healthcare experience (mean= 11.573, SD = 1.039) were statistically different from each other. The analysis determined the results were not statistically significantly different (t (33) = -0.402, p = 0.69). The full results are featured in Table 12.

Table 12 - 2014 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without Healthcare Experience

Healthcare Experience	N	Mean MMI Score	Std. Deviation	Std. Error Mean
No	5 (14%)	11.573	1.039	0.465
Yes	30 (86%)	11.848	1.456	0.266

3.4b. 2014 Candidate Professional Work Experience

An independent sample t-test compared MMI score to candidates with and without professional work experience. While those with professional work experience had a higher mean MMI score (12.047), than those without (11.525), the t-test results were not significantly statistically different (t (33) = -1.108, p = 0.276). Tables 13 illustrates these results.

Table 13 - 2014 Mean, SD and Standard Error of the Mean for candidate's MMI performance with and without professional work experience

Professional Work Experience	N	Mean	Std. Deviation	Std. Error Mean
No	16 (46%)	11.525	1.433	0.358
Yes	19 (54%)	12.047	1.352	0.310

3.4c. 2014 Candidate Experience with children with disabilities

An independent sample t-test was conducted to compare MMI score to candidate's who had experience with children with disabilities (mean =12.097, SD =1.322), and those without (mean = 11.693, SD = 1.432). The analysis determined the results were not statistically significantly different (t (33) = -0.768, p = 0.448). The full results are featured in Table 14.

Table 14 - 2014 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without experience with children with disabilities

Experience with disabilities	N	Mean	Std. Deviation	Std. Error Mean
No	25 (71%)	11.693	1.432	0.286
Yes	10 (29%)	12.097	1.322	0.418

3.4d. 2014 Candidate Previous Child Life Course Work

The results of the independent sample t-test are featured in Table 15. Those who had previously taken child life course work had a mean MMI score of 12.02, whereas those who had not taken such course work had a mean MMI score of 11.724. The analysis

determined the results were not statistically significantly different (t (33) = -0.562, p = 0.578).

Table 15 - 2014 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without previous child life course work

Previous Child Life Course work	N	Mean	Std. Deviation	Std. Error Mean
No	25 (71%)	11.724	1.489	0.298
Yes	10 (29%)	12.020	1.166	0.369

3.4e. 2014 Candidate Undergraduate Degree Type

Table 16 identifies the descriptives for previous undergraduate degree type and its relationship to candidate's overall MMI scores. . There were a total of seven different degree types among the 35 interviewed candidates in 2014. A total of 42.8% (15/35) of candidates came from child development related degrees. There was a fairly even spread of remaining degree types. The same number of candidates (11% or 4/35) held degrees in either science, psychology or in the social sciences respectively. Eight and a half percent (3/35) of candidates had a degree in health or in the humanities, and 5.7% (2/35) had a degree in the category identified as 'other'.

Table 16- 2014 Descriptives Related to Interviewed Candidate's Undergraduate Degree Type and Overall MMI Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min.	Max.	Between-Component Variance
					Lower Bound	Upper Bound			
Health degree	3	12.944	1.781	1.029	8.518	17.370	10.900	14.167	
Science degree	4	12.542	0.495	0.247	11.755	13.329	11.933	13.133	
Child Dev. Related Degree	15	11.036	1.449	0.374	11.755	11.838	8.467	14.133	
Psychology Related degree	4	11.708	1.444	0.722	9.411	14.006	10.433	13.700	
Social Sciences	4	12.383	1.092	0.722	10.646	14.121	10.900	13.433	
Humanities	3	12.111	0.139	0.080	11.766	12.456	12.000	12.267	
Other	2	13.033	0.330	0.233	10.069	15.998	12.800	13.267	
Total	35	11.809	1.39 4	0.236	11.330	12.288	8.467	14.167	
Fixed Effects			1.289	0.218	11.363	12.255			
Random Effects				0.367	10.910	12.708			0.364

Health degree includes; Bachelor of Health Sciences, Child Health
Science degree includes; Medical Science, Biology, Microbiology, Kinesiology, Phys. Ed.
Child Development Related degree includes; Child & Youth Studies, Child, Youth & Family, Early Childhood Studies, Childhood & Social Institutions, Child Development, Contemporary Studies
Psychology Related degree includes; Psychology, Life Sciences
Social Sciences degree includes; Social Work, Social Services, Sociology, Social Sciences, Political Sciences
Humanities degree includes; Fine Arts, English Communications, Philosophy, History
Other degree includes; Therapeutic Recreation, Engineering

Table 17 shows the results of the one-way ANOVA of degree type ($F_{(6, 28)}$ = 1.970, p = 0.104).

Table 17 - 2014 One-Way ANOVA of Undergraduate Degree Type

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	19.620	6	3.270	1.970	0.104
Within Groups	46.472	28	1.660		
Total	66.092	34			

3.5 Part B 2015 Reliability Results

The 2015 G-study analysis with UR Genova yielded overall test generalizability equal to 0.089. Table 18 reports these results and the variance components. Minimal variation is found among station (0.046); question nested within station (0.046) and candidates (0.006).

Table 18 -2015 Summary of estimated Variance Components, effects and the G-coefficient (Experiment 2 - 2015)

Facet	Df	MS	Variance
Candidate	39	0.741	0.006
Station	3	6.311	0.046
Question nested in Station	7	0.852	0.013
G-coefficient	0.089		

A D Study was similarly performed for 2015 to determine if alterations to the number of observations would improve levels of reliability. The results are reported in Table 19 and include what might be expected of an 8-station (0.171) and 12-station MMI

(0.237) in addition to what might be expected if station 4 were modified to include 3 questions (0.094).

Table 19 D Study - 2015 Predicted reliability if increased stations

Number of Stations	Number of Questions per Station	D-Study Adjusted Co-efficient
4	3	0.094
8	3	0.171
12	3	0.237

The range of scores, mean and standard deviation for the MMI questions and station averages are included in Table 20. The minimum score per question achieved is 1 and the maximum is 4. Station 1 yields the highest mean scores with the station average at 3.35. Stations 2 and 3 have very similar station averages with 2.908 and 2.975 respectively. The Station 4 average is 2.793.

Table 20 - 2015 Range of Scores, Mean & Standard Deviation for MMI Questions & Station Averages

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Station 1 Question 1	40	2	2	4	3.400	0.709
Station 1 Question 2	40	2	2	4	3.480	0.554
Station 1 Question 3	40	2	2	4	3.180	0.712
Station 1 Average	40	1.670	2.330	4	3.350	0.584
Station 2 Question 1	40	3	1	4	2.780	0.768
Station 2 Question 2	40	3	1	4	2.950	0.846
Station 2 Question 3	40	2	2	4	3.00	0.716
Station2 Average	40	2.330	1.670	4	2.908	0.560
Station 3 Question 1	40	2	2	4	2.900	0.591
Station 3 Question 2	40	3	1	4	2.950	0.597

Station 3 Question 3	40	3	1	4	3.080	0.656
Station3 Average	40	2	2	4	2.975	0.373
Station 4 Question 1	40	2.400	1.400	3.800	2.960	0.511
Station 4 Question 4	40	2	2	4	2.630	0.540
Station 4 Average	40	2.100	1.700	3.800	2.793	0.417

The mean and standard deviation of question scores and station averages were also

completed to compare scores between the two different interview days held in 2015. The

results are reported below in Table 21.

Table 21 - 2015 Mean & Standard Deviation for MMI Questions & Station per Interview Day

	Day 1 Mean	Day 1 Std. Deviation	Day 2 Mean	Day 2 Std. Deviation
Station 1 Question 1	3.4	0.82	3.4	0.6
Station 1 Question 2	3.4	0.6	3.55	0.51
Station 1 Question 3	3.1	0.79	3.25	0.64
Station 1 Average	3.3	0.639	3.4	0.536
Station 2 Question 1	2.95	0.76	2.6	0.75
Station 2 Question 2	3.15	0.88	2.75	0.79
Station 2 Question 3	3.25	0.64	2.75	0.72
Station 2 Average	3.117	0.475	2.7	0.571
Station 3 Question 1	3.05	0.6	2.75	0.55
Station 3 Question 2	3.1	0.6	2.8	0.6
Station 3 Question 3	3.1	0.64	3.05	0.69
Station 3 Average	3.083	0.373	2.867	0.349
Station 4 Question 1	2.83	0.544	3.09	0.452
Station 4 Question 4	2.6	0.503	2.65	0.587
Station 4 Average	2.715	0.449	2.87	0.379

Pearson correlations were also used to assess the relationship between overall station

scores. Results of the 2015 data are shown in Table 22. There are no correlations

between station average scores.

Table 22 Pearson Correlations for 2015 MMI Station Average Scores (n=40)

2015		Station 1 Average	Station 2 Average	Station 3 Average	Station 4 Average
Station 1 Average	Pearson Correlation	1	-0.047	-0.050	0.102
	Sig. (2-tailed)	■	0.771	0.758	0.530
Station 2 Average	Pearson Correlation	-0.047	1	0.153	-0.003
	Sig. (2-tailed)	0.771	■	0.347	0.985
Station 3 Average	Pearson Correlation	-0.050	0.153	1	-0.040
	Sig. (2-tailed)	0.758	0.347	■	0.808
Station 4 Average	Pearson Correlation	0.102	-0.003	-0.040	1
	Sig. (2-tailed)	0.530	0.985	0.808	■

** Correlation is significant at the 0.01 level (2-tailed)

* Correlation is significant at the 0.05 level (2-tailed)

3.6 2015 Predictors of MMI Performance

The applications of the 40 interviewed candidates in 2015 were also reviewed to collect their background information and discern the relative homogeneity or heterogeneity of the pool of interviewed candidates. This information was also used to run 4 separate independent t-tests to see if there were any predictors of candidate performance on the MMI based on the same predictors; experience in health care; professional experience; experience with children with disabilities; and having taken

previous child life course work from this program's online undergraduate elective options. These t-test results will be reported in the sequence that predictors were identified above.

3.6a. 2015 Candidate Healthcare Experience

An independent sample t-test was conducted to compare MMI score to candidate's with experience in healthcare (mean= 12.016, SD = 1.047) and without healthcare experience (mean= 12.057, SD = 0.93). The analysis determined the results were not statistically significantly different (t (38) = 0.11, p = 0.91). The full results are listed in Table 23.

Table 23 - 2015 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without Healthcare Experience

Healthcare Experience	N	Mean	Std. Deviation	Std. Error Mean
No	10 (25%)	12.057	0.930	0.294
Yes	30 (75%)	12.016	1.047	0.191

3.6b. 2015 Candidate Professional Work Experience

An independent sample t-test was used to compare MMI score to candidates with and without professional work experience. Those with professional work experience had a mean MMI score of 12.347, and SD of 0.963. Those without had a mean score of 11.888, and SD of 1.011. The analysis determined the results were not statistically significantly different (t (38) = -1.33, p = 0.19). Table 24 illustrate these results.

Table 24 - 2015 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without Professional Work Experience

Professional Work Experience	N	Mean	Std. Deviation	Std. Error Mean
No	28 (70%)	11.888	1.011	0.191
Yes	12 (30%)	12.347	0.963	0.278

3.6c. 2015 Candidate Experience with children with disabilities

Table 25 reports the results of the independent sample t-test conducted to compare MMI score to candidate's who had experience with children with disabilities (mean =12.175, SD =1.066), and those who did not (mean = 11.916, SD = 0.971). The analysis determined the results were not statistically significantly different (t (38) = -0.799, p = 0.429).

Table 25 - 2015 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without experience with children with disabilities

Experience with Children with Disabilities	N	Mean	Std. Deviation	Std. Error Mean
No	23 (57.5%)	11.916	0.971	0.203
Yes	17 (42.5%)	12.175	1.066	0.258

3.6d. 2015 Candidate Previous Child Life Course Work

An independent sample t-test was used to compare MMI scores to those who had previously taken child life course work, and those who had not. Those with previous child life course work had a mean MMI score of 12.30 and SD of 0.536, and those who had not

taken such course work had a mean MMI score of 11.894 and SD of 1.155. The analysis

determined the results were not statistically significantly different (t (38) = -1.201, p =

0.237). The results of the independent sample t-test are featured in Table 26.

Table 26 - 2015 Mean, SD and Standard Error of the Mean for Candidate's MMI Performance with and without previous child life course work

Previous Child Life Course Work	N	Mean	Std. Deviation	Std. Error Mean
No	27 (67.5%)	11.894	1.155	0.222
Yes	13 (32.5%)	12.300	0.536	0.149

3.6e. 2015 Candidate Undergraduate degree types

Table 27 identifies the descriptives for previous undergraduate degree type and its

relationship to candidate's overall MMI scores in 2015. There were a total of seven

different degree types among the 40 interviewed candidates. The three most common

degree types among the pool of interviewed candidates were those in health (27.5% or

11/40), child development related degrees (27.5% or 11/40) and psychology related

degrees (25% or 10/40). Eighty percent (32/40) of candidate's interviewed held these

degrees.

Table 27 - 2015 Descriptives Related to Interviewed Candidate's Undergraduate Degree Type and Overall MMI Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min.	Max.	Between-Component Variance
					Lower Bound	Upper Bound			
Health degree	11	12.697	1.068	0.322	11.979	13.415	10.400	14.233	
Science degree	3	11.944	0.847	0.322	9.841	14.048	10.967	12.433	
Child Dev. Related Degree	11	11.545	0.799	0.241	11.009	12.082	10.467	12.667	
Psychology Related degree	10	11.853	1.016	0.321	11.127	12.580	9.767	13.300	
Social Sciences	2	12.233	0.754	0.533	5.457	19.010	11.700	12.767	
Humanities	3	12.233	1.117	0.645	9.070	14.619	10.567	12.633	
Total	40	12.026	1.007	0.159	11.704	12.348	9.767	14.233	
Fixed Effects			0.964	0.152	11.716	12.336			
Random Effects				0.219	11.463	12.588			0.108

Health degree includes; Bachelor of Health Sciences, Child Health
Science degree includes; Medical Science, Biology, Microbiology, Kinesiology, Phys. Ed.
Child Development Related degree includes; Child & Youth Studies, Child, Youth & Family, Early Childhood Studies, Childhood & Social Institutions, Child Development, Contemporary Studies
Psychology Related degree includes; Psychology, Life Sciences
Social Sciences degree includes; Social Work, Social Services, Sociology, Social Sciences, Political Sciences
Humanities degree includes; Fine Arts, English Communications, Philosophy, History
Other degree includes; Therapeutic Recreation, Engineering

Table 28 shows the results of the one-way ANOVA for degree type (F (5, 34) = 1.722, p = 0.156).

Table 28 - 2015 One-way ANOVAs of Degree Type

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7.995	5	1.599	1.722	0.156
Within Groups	31.577	34	0.929		
Total	39.572	39			

CHAPTER 4: DISCUSSION

This thesis sought to understand the following study questions;

1. Is a 4-station MMI feasible for small programs to undertake?

2. Is a 4-station MMI acceptable among child life candidates and interviewers?

3. Is this 4-station MMI design reliable?

4. Are there any predictors of candidate's MMI performance?

5. Are there pitfalls in small MMIs that can be minimized or avoided?

The following discussion will provide an interpretation of the study findings to address these questions in sequential order.

4.1 Feasibility of MMI for small programs

This study sought to determine if a 4-station MMI was feasible for small programs to utilize in high-stakes selection. Interviews are known to be a costly undertaking for all academic programs. Regardless of program size, the burden of cost has been so significant that some institutions have chosen to discontinue admissions interviews altogether (Gabard, et al., 1997). In the health professions, knowledge is critical in delivering efficacious patient care, but the manner in which that care is delivered is similarly important for the human experience and the development of trust within the care provider and patient relationship. It is however difficult to measure personal and professional characteristics utilized in the delivery of patient care (e.g. communication, empathy) with only GPA.

Concern has been raised in the literature about the costs associated with transitioning from panel interviews to the MMI format (Axelson & Kreiter, 2009). Through a comparison of this program's experience with both formats, there were more costs up front in developing and implementing MMI, but there were cost savings to be had in the end in staff time. The panel interview format previously required 4 full interviewing days to interview and score 24 candidates for a total of 120 combined staff hours. Given the part-time appointments of faculty members and their other positions outside the university, interviewers often took vacation time from their other roles to interview because of the number of days and dedicated time needed.

Similar to other studies assessing feasibility of MMI interviews, with the implementation of MMI, a larger number of candidates could be interviewed in a shorter time (O'Brien, Harvey, Shannon, Lewis, Valencia, 2011; Knorr & Hissbach, 2014) and was found considerably more efficient in the number of staff hours needed to evaluate candidates (Rosenfeld, Reiter, Trinh, Eva, 2008).

Additional staff time was however needed for MMI question development, but this total time was not tracked. The amount of staff time needed was most dependent on the type of question(s) being developed. While Eva & Macala (2014) reported, behavioural interview questions were found to be the least time consuming and easiest to generate, this program found the writing component question within station 4 took the least amount of time to prepare, followed by behavioural interviewing questions or experienced based questions, with the situational-based questions, where candidates imagine what they would do in that situation, taking the longest.

In the previous panel interviews utilized, each candidate could use the time remaining within their interview to ask questions about the program, university, clinical education, among others. While the specific number of inquiries received and hours spent addressing them were not tracked, a greater amount of staff time was spent addressing correspondence received post-personal interviews from candidates with questions about the program, as some had limited remaining time to ask questions, and some did not choose to ask any questions at that time in the immediacy of just completing their panel interview. A noticeable reduction in staff time was however found addressing candidate inquiries with the initiation of the standard length question and answer period held following the MMI interviews. Although there was varied time candidates had to process their MMI experience before this question and answer period, this format provided all candidates with an equal amount of time and opportunity to ask questions, build on the questions of others, and receive information.

4.1 Part A 2014 Findings

During the 2014 MMI schedule, a total of 36 candidates could be interviewed, although one did not attend. In this staggered start design, a total of 72.75 hours of staff and volunteer time were required over three days. This included a total of four staff/faculty and one volunteer to run the check-in desk.

More staff time was however needed for development of MMI questions, and station scoring sheets, particularly in this inaugural year. This time was not formally tracked. A greater number of situation-based questions were used in 2014, followed by behavioural interviewing questions which were then added to the program's question

bank. Selecting an appropriate abstract for the writing question in station 4 took the least amount of time. More staff time was also required to coordinate multiple station/interview spaces in close proximity, although this time was also not formally tracked.

4.1 Part B 2015 Findings

The modified 2015 MMI schedule enabled interviewing and evaluating 40 candidates in just two interview days. Candidates were pre-assigned into one of five groups of four during their interview day. A total of 57 hours of staff and 36 hours of volunteer time were utilized for a total of 93 hours over the two days. Unlike 2014, candidates did not progress through the stations in the same sequential order, but in the order they were assigned.

Similar to the findings of Rosenfeld, et al., 2008, more support staff were required to be in attendance during the MMIs. With the increase in candidate numbers on the MMI interview days, more manpower was needed to facilitate the flow of the MMI process than necessary for panel interviews. Fortunately, new graduates were eager to participate as volunteers. While more interview room space was needed within close proximity, the shortened span of days needed enabled use of tutorial rooms on the weekend when they were in least demand. While there were no costs for use of these rooms during the MMIs within this study, housekeeping and security costs may be needed for future weekend use of tutorial room space.

More preparation time was however needed in developing new MMI questions in comparison to previous practice where minor revisions would occur yearly to the panel

interview tool prior to its use. One of the primary hurdles to the adoption of the MMI process noted in the literature has in fact related to the creation of structured stations (Rosenfeld, et al., 2008). More time was spent on question development in 2014, with less time utilized in 2015 as more behavioural interviewing questions were included from a time efficiency perspective. Discussion on whether this time savings was evaluated to provide a cost-benefit will be continued in the reliability section ahead. This initial burden of question preparation is however expected to decrease as the bank of questions grows with each MMI cycle (Rosenfeld, et al., 2008).

4.2 Acceptability of MMI Among Child Life Candidates and Faculty Interviewers

There was a positive response to the use of MMI in applicant selection by both candidates and interviewers similar to the findings of Eva, et al., 2004a and Humphrey, et al., (2008). All candidates agreed the interview process was well organized, enjoyable (97.5%), and provided them sufficient time to present their ideas (92.5%). A total of 97.5% of candidates preferred meeting 1 on 1 with an interviewer in each station instead of with a traditional interview panel, and indicated they were satisfied with the MMI being used as one of the tools for admissions selection (92.5%). Compared to traditional panel interviews, 82.5% agreed the MMI was more fair.

One area that rated lower among survey responses was candidate's level of agreement that they could accurately portray themselves during the MMI. Eighty percent rated this question as strongly agree or agree, while 15% agreed somewhat, 2.5% neither agreed nor disagreed and 2.5% disagreed somewhat. This is similar to the qualitative study by Kumar, et al. (2009) where candidates described limited opportunities to present holistic

pictures of themselves. Even though a purposeful attempt was made in the MMI design to include a question for candidates to self-present on why they should be chosen for the program, this still rated lowest among this survey's questions, and rated lower than respondents to a similar candidate survey question asked in the study by Dore, et al., (2010). Eight-eight percent of those survey respondents believed they could accurately portray themselves.

All four interviewers had experience with the previously used traditional panel interview format. There was 100% agreement that their workload was less with the MMI; found the MMI to be more fair; preferred meeting 1-1 with candidates instead of the panel format; and would be in favour of continuing its use instead of traditional interviews. While these findings are similar to those in other studies (Fraga, et. al., 2013), the interviewers within this study rated these as 7 out of 7 on the Likert scale provided.

One area which rated lower in interviewer survey scores was in the question which asked about their level of agreement that the scoring sheet allowed them to differentiate among applicants. One of the four interviewers indicated that she disagreed somewhat as she thought a broader scale (e.g. 7-point scale) may be better to use than the current 4-point scale. The remaining three however strongly agreed that the 4-point scale utilized allowed them to differentiate among candidates.

Similar to the qualitative comments made in the Kumar et al., 2009 study, open text responses of the interviewers similarly reported that the 1-1 interview format appeared to improve the quality of the interactions with candidates, and reduced

candidate stress. Study findings therefore indicate that a 4-station MMI is highly acceptable among child life candidates and interviewers in small academic programs.

There was also a positive response to the post MMI Question and Answer period with 100% of candidates either agreeing or strongly agreeing that it was informative, addressed their questions and provided them with greater knowledge about the Child Life Studies Program and faculty. Inclusion of this component in the MMI schedule therefore addressed a gap noted by others in understanding program personality (Phillips & Garmel, 2014) and also minimized the number of emails historically received by individual candidates with program related questions when using the panel interview format. As noted by the open text responses within the survey, the question and answer period was perceived as helpful as candidates indicated it was "beneficial to be with others who asked questions which created more questions and more learning".

4.3 Part A - Reliability of the 2014 4-station MMI design

This study also sought to understand if this 4 station MMI design was reliable in the context of a Child Life Study program selection. In 2014, the G-study analysis of the 4-station MMI yielded an acceptable level of reliability at 0.718. This is similar to others reported in the literature with 7 to 12 station MMI designs (Dore, et al, 2010; Roberts, et al., 2008; Tiller, et al., 2013; Uijtdehaage, et al., 2011; Hofmeister, et al., 2009). The D-study results show this can be improved even further with one minor station design change, while maintaining 4-stations. Three stations within this MMI had 3 questions per station. Station 4 however had only 2 due to the written component. By modifying this station to include 3 questions, it is predicted to improve reliability to 0.734. The D-study

revealed that additional alterations to the number of stations and questions collected can further improve levels reliability. By increasing to 8 or 12 stations with 3 questions (24 - 36 questions therefore needed) reliability is predicted to increase to 0.847 - 0.892. While reaching this level of reliability is very appealing, the feasibility to undertake is far too costly to consider, particularly when a reasonable level of reliability has already been achieved.

The correlation of MMI stations in 2014 was an unexpected finding. The cause of this remains unclear. Station 1 correlated with all other stations. Replacing the easier child life profession specific questions within this station with more challenging scenarios, is hypothesized to reduce the level of station correlations.

Part B - Reliability of the 2015 4-station MMI design

The G-study analysis of the 4-station MMI conducted in 2015 however yielded an unexpected reliability. The overall test generalizability was equal to 0.089, considerably lower than the previous year. Increasing the stations through the D-study did not improve reliability to an acceptable level. The 4-station design, number of questions used, and interviewers however remained the same. Two variables however changed between years; the pool of candidates interviewed, and the types of questions used.

While there has always been a very heterogeneous group of applicants to the Child Life Studies Program, a review of interviewed candidate's backgrounds revealed that those applicants offered interviews in 2015, were very similar in different ways than those in the 2014 pool of interviewed candidates. The variance component attributable to

candidate differences was found to only be 0.006, "although in most studies candidates accounted for <30% of the variance" (Knorr & Hissbach, 2014, p.1163). While both groups in 2014 and 2015 had high numbers of individuals with healthcare experience (87.5% (30/35) and 75% (30/40) respectively), almost half (42.5% (17/40)) of the group in 2015 had experience with children with disabilities, and had taken child life coursework offered by this program as undergraduate electives. This increased their clinical and academic exposure to child life. In addition, there was greater similarity in undergraduate degree type (3 degree types accounted for 87.5% (32/40) of those interviewed) which further tightened the cluster of applicants interviewed making detection of differences among candidates difficult.

As previously noted the types of questions also changed in the 2015 MMI design. Unlike 2014, only one situational question was used. Station 1 maintained its focus on an applicant's understanding of the role of a child life specialist in healthcare, and station 4 also remained similar in both years, with only a change in the research abstract used for the writing component. Stations 2 and 3 however changed to only include behavioural interviewing questions.

Behavioural interview or experienced based questions where candidates are asked to "provide an example of a time when...", or "describe a situation where...", may have made it easier for candidates to rehearse or prepare for these kinds of questions. Kumar and colleagues (2009) report that scenario based MMI can make it more difficult for rehearsal and coaching of responses. Prior exposure to scenario based MMI station stems have in fact been found to not have an impact on MMI performance ratings assigned (Reiter,

Salvatori, Rosenfeld, Trinh, Eva, 2006). In addition, situational judgement tests, or low

fidelity simulations, have been found to be valid predictors of later performance in high-

stakes settings (Patterson, Lievens, Kerrin, Zibarras, Carette, 2012). In contrast,

behavioural questions may be easier to recall and share in discussion. Interview groups in

2015 were not sequestered, and the same questions were used during both interview days.

While it cannot be ruled out that behavioural questions were shared so there was time for

other candidates to rehearse and/or prepare for their responses, the comparison of day one

and two question and station averages do not show a significant difference in candidate's

performance on day two of the interviews. The station averages in 2015 were very similar

in station one and four, and the averages were in fact less on day two in stations 2 and 3.

This is similar to the comparative findings in 2014 when more scenario based questions

were used across 3 days.

While the use of a high number of behavioural interviewing questions was useful from

a feasibility perspective in 2015, as this question type can be developed in less time, such

a high use of them does appear to have had an impact on reliability. Gafni and colleagues

found in their study, that increasing the number of behavioural stations from six or seven

to fourteen or fifteen provided an increase of 0.02 to 0.15 in the reliability coefficient

(Gafni, Moshinsky, Eisenberg, Zeigler & Ziv, 2012). The small number of stations within

this study design, and high use of behavioural questions in 2015, may have therefore led

to reduced reliability.

A study by Eva and Macala (2014) sought to investigate the types of structure built into

question stations with individuals (n=41) invited to interview in their university's

undergraduate medical program. Candidates were eligible to participate only after completing the regular medical school admissions interview. In this study's design, four stations were presented in three different ways. These different ways included behavioural interviewing, where applicants were asked to recall from experience what they did in specific situations; situational judgements where asked to imagine a scenario and what they would do; and free form stations where they were only given a brief explanation about what the station was intended for and no further information on how to carry out the discussion. In contrast, their findings suggest that behavioural interviewing was more reliable in differentiating among candidates (G= 0.77) in comparison to situational judgement types (G=0.69) and free form stations (G=0.66). Given that this is the only study of its kind, and there only being moderate differences in reliability between question types, this warrants further analysis on whether there is any difference in reliability for high stakes decision making, particularly when varying numbers of MMI stations are utilized.

4.4 Predictors of candidate's MMI performance

Healthcare experience, professional work experience, experience with children with disabilities, previous child life course work and degree types were hypothesized to predict higher MMI performance. None of these experiences however resulted in statistically significant findings as all values were greater than 0.05 in both 2014 and 2015. While there may have been slight differences between mean scores, the prevalence of these predictor variables among candidates, particularly in 2015, made the candidate groups more homogenous.

4.5 Pitfalls in small MMI to Minimize or Avoid

This study demonstrates that there can be both success and challenges in running a 4-station MMI. As a small allied health program, the goal has been to increase the reliability of our admissions process. However, through data analyses, several pitfalls have been identified. These will be discussed along with the next steps planned to address them with generalizability to other professions and admissions teams.

Eva and Macala (2014) identify that "moderate reliability can be achieved simply by aggregating across many observations" (p. 611). This was found to be true with our 2014 procedures, but not the case in 2015. This demonstrates that an assessment measure that inadequately differentiates among applicants cannot always be strengthened with increased observations alone.

The Pearson's correlation data suggests that while our intent was good to ensure candidates knew about the child life profession and what they would be signing up for should they be offered admission into the program, a potential ceiling effect occurred in Station 1 where these questions were asked. Having a homogenous group of applicants with similar experiences, incoming knowledge about the profession, and perhaps even opportunity to rehearse in their interview preparations for such questions made this station too easy to allow for differentiation of candidates according to ability. Care must be taken in creating station questions that are more difficult for candidates to prepare for through coaching and rehearsal of responses (Kumar, et al., 2009). Out of a maximum score of 4, the mean scores for this station were 3.5429 in 2014 and 3.35 in 2015. "To facilitate discrimination between applicants, the stations must have an optimal level of

difficulty" (Uijtdehaage et al., 2011). Having an entire station devoted to questions specific to the profession (3 out of a total of 11 questions) did not therefore provide value in screening in and out candidates. It is recommended that similar pitfalls be avoided in MMI development with careful consideration given to the ratio of questions asked on profession-specific topics. Per these findings, a single question would have been sufficient to attend to the faculty's interest in addressing candidate's knowledge about the profession, while providing opportunity to ask two other questions that are less easy to adjust to optimal level of station difficulty and better differentiate candidates.

Feasibility will always remain an important consideration for any undertaking. The reliability of the MMI increases with the number of stations, and the number of stations depends on individual program resource availability (Dore et. al., 2010, p.S62). It therefore remains important to pick and choose what has the least chance of effecting your tools acceptability and reliability if cost is a factor. While the inclusion of a writing station saved some of the costs in designing a new situational scenario for example, other studies have found that this component does not increase reliability (Raghaven, Burnett, Martin, Christensen, Young, Mackalski, Aoki, 2013).

In this particular study design, the writing component unfortunately took up the time of 2 questions for candidates to complete, but provided only one score. This led to a reduction in the number of more reliable observations, and also created disparity in the number of questions per station. The D-study analyses demonstrated an increase in reliability by modifying this station to align with the others, and include three questions per station. Removing the writing component therefore addresses two pitfalls. It would

enable the addition of more reliable question formats to increase the overall reliability of the candidate assessment tool, and provide consistency in the number of questions asked per station.

There is the possibility that there could be volatility of this small MMI station design between years. An increase in stations could be considered while utilizing a similar number of raters and time resources through a double circuit whereby candidates could do two cycles through the same 4 raters. Instead of having15 minute stations with three questions and five minutes to score, a model could be used where interview stations are eight minutes in length and two minutes to score. Candidates could complete stations 1-4 in the first 40 minutes, and stations 5-8 in the next 40 minutes, with the rater in the first room doing stations 1 & 5, the rater in the second doing stations 2 & 6, and so on thereby enabling an increase in the number of independent observations. In an unpublished study by Rosenfeld and Eva, they tracked how many candidates examiners saw in between repeat performances. Through personal communication with Kevin Eva, it was reported that the "correlation between successive observations of a given candidate dropped from 0.5 if 0 or 2 interactions intervened to 0.3 if >2 interactions intervened". With each station 8 minutes in length with an additional two minutes for scoring, this means 20 minutes passed from the time any rater saw an applicant the first to the second time in the double circuit to get to the r=0.3. Eva and Rosenfeld (personal communication) concluded that there was no absolute influence of seeing an examiner for the second time in terms of mean score. The amount of time/number of stations that occurred prior to examiners seeing a candidate the second time, and station order did not show statistical difference in

mean scores. The individual scores given on the questions within each station within this 4-station design were exceptionally close. The double circuit methodology therefore becomes a strong option to both increase independent observations, and extinguish any halo effect which may have occurred when several questions were asked within stations.

The use of MMI holds promise for other small programs. Even with all the pitfalls identified within this study, the MMI still compares favourably to the traditional panel interview. As previously discussed, the MMI station format is strengthened by repeated sampling of performance to address context specificity, and reduce situational and interviewer biases for which traditional panel interviews are noted for. The continued use of MMIs is further supported through the systematic review findings of Patterson and colleagues (2016) on the relative strengths of selection methods. Their findings show that structured interviews or MMIs, situational judgement tests and selection centres "are more effective methods and generally fairer than traditional interviews, references and personal statements" (p.48). The levels of acceptability demonstrated within this study also show how well received its use is among this population of stakeholders. This, in combination with the increase in the number of candidates that can be interviewed in a shorter period of time, and the enhanced reliability that can occur when the pitfalls are minimized, make this tool a worthy endeavour to pursue.

5. STUDY STRENGTHS AND LIMITATIONS

This is the first MMI used in child life candidate selection and among the psychosocial professions. While the sample sizes are somewhat small, this study has demonstrated high acceptability for the use of MMI among its candidates (n=40) and

interviewers (n=4). Given that stakeholder surveys were collected immediately after the MMIs and before the results of the interviews were known, as in the Kumar et al. (2009) study, this added to the authenticity of the data collected. There is however the possibility that collecting this information right after could have affected responses due to candidate and/or interviewer fatigue, and their lack of time to reflect on the process. Given there is no other study to evaluate the acceptability of MMI in child life candidate selection, this will add to the literature base with generalizability to other small professional programs.

This study has also sought to analyze the differences in this specific MMI design between cohort years as there was a notable change in levels of reliability. As Knorr et al. (2014) report, analyzing how systematic changes influence reliability in a specific MMI design is of high value. The pitfalls identified here and specific suggestions provided to address them will provide other programs with important considerations to avoid or minimize the same challenges.

6. FUTURE DIRECTIONS

The article by Knorr et al., (2014) identifies that while there may be a number of studies addressing MMI and academic institutions implementing MMI, there are a variety of approaches in how each one chooses to carry them out. It can therefore be difficult to separate the variance components and compare proportions of variance between studies (Knorr, et. al, 2014). Adding more to the literature base on smaller scaled MMIs like this one provides more specific comparators for others creating MMIs similar in sample size, and station number. While it can be difficult for smaller programs to consider developing

an 8-12 station MMI with limited resources, the insight gathered here may inform others outside of child life on the process to implement a smaller station design, or a double circuit with the same number of raters, particularly with the data available on the feasibility to implement.

It is clear from this study that there is more to be explored in terms of station type. With many different station formats reported throughout the MMI literature, understanding the reliability of each format (e.g. behavioural interviewing, situational scenarios) and if there are measurable differences between them in terms of reliability, will provide all programs with critical information to use when developing the focus of each of their stations, and in determining the ratio of different formats to include, for incremental validation of their MMI process.

APPENDIX A: Child Life Studies Program Panel Interview Schedule 1990 - 2013

6 candidates per day x 4 days = 24 candidates interviewed

9:00am - 10:00am	Interview Candidate 1
10:00am - 10:15am	Score interview
10:15am -11:15am	Interview Candidate 2
11:15am - 11:30am	Score interview
11:30am - 12:30pm	Interview Candidate 3
12:30pm - 12:45pm	Score interview
12:45pm - 1:30pm	Lunch
1:30pm - 2:30pm	Interview Candidate 4
2:30pm-2:45pm	Score interview
2:45pm - 3:45pm	Interview Candidate 5
3:45pm - 4:00pm	Score interview
4:00pm - 5:00pm	Interview Candidate 6
5:00pm - 5:15pm	Score interview

* Of note, until the early 2000s, a group interview was held each evening.

APPENDIX B: Child Life Studies Program 2014 Multiple-Mini Interview Schedule

12 candidates per day x 3 days = 36 candidates interviewed

Time	Interviewer 1	Interviewer 2	Interviewer 3	Interviewer 4
9:00am - 9:15am	Candidate 1			
9:20am - 9:35 am	Candidate 2	Candidate 1		
9:40am – 9:55am	Candidate 3	Candidate 2	Candidate 1	
10:00am –10:15am	Candidate 4	Candidate 3	Candidate 2	Candidate 1
10:20am - 10:35am	Candidate 5	Candidate 4	Candidate 3	Candidate 2
10:40am –10:55am	Candidate 6	Candidate 5	Candidate 4	Candidate 3
11:00am –11:15am		Candidate 6	Candidate 5	Candidate 4
11:20am –11:35am			Candidate 6	Candidate 5
11:40am –11:55am				Candidate 6

12:15 -12:45pm   Program Info & Question Period with interviewers 1 & 2

Time	Interviewer 1	Interviewer 2	Interviewer 3	Interviewer 4
1:00pm - 1:15pm	Candidate 7			
1:20pm – 1:35pm	Candidate 8	Candidate 7		
1:40pm – 1:55pm	Candidate 9	Candidate 8	Candidate 7	
2:00pm – 2:15pm	Candidate 10	Candidate 9	Candidate 8	Candidate 7
2:20pm – 2:35pm	Candidate 11	Candidate 10	Candidate 9	Candidate 8
2:40pm – 2:55pm	Candidate 12	Candidate 11	Candidate 10	Candidate 9
3:00pm – 3:15pm		Candidate 12	Candidate 11	Candidate 10
3:20pm – 3:35pm			Candidate 12	Candidate 11
3:40pm - 3:55pm				Candidate 12

4:15pm – 4:45pm   Program Info & Question Period with interviewers 3 & 4

APPENDIX C: Child Life Studies Program 2015 Multiple-Mini Interview Schedule

20 candidates per day x 2 days = 40 candidates interviewed

Round 1 – 9 am – 10:15 am

Event	Timing	Candidate 1	Candidate 2	Candidate 3	Candidate 4
Candidate Check-in	8:30 am				
Volunteer Leads Candidates to Assigned Stations	8:50 am				
Enter Station	9 am	A	B	C	D
Exit Station	9:15 am				
Enter Next Station	9:20 am	B	C	D	A
Exit Station	9:35 am				
Enter Next Station	9:40 am	C	D	A	B
Exit Station	9:55 am				
Enter Next Station	10 am	D	A	B	C
Exit Station Volunteer Leads Candidates room to complete exit survey	10:15 am				

Round 2 – 10:20 am – 11:35 am

Event	Timing	Candidate 5	Candidate 6	Candidate 7	Candidate 8
Candidate Check-in	10 am				
Volunteer Leads Candidates to Assigned Stations	10:15 am				
Enter Station	10:20 am	A	B	C	D
Exit Station	10:35 am				
Enter Next Station	10:40 am	B	C	D	A
Exit Station	10:55 am				
Enter Next Station	11 am	C	D	A	B
Exit Station	11:15 am				
Enter Next Station	11:20 am	D	A	B	C
Exit Station Volunteer Leads Candidates to room to complete exit survey	11:35 am				

Round 3 – 11:40 am – 12:55 pm

Event	Timing	Candidate 9	Candidate 10	Candidate 11	Candidate 12
Candidate Check-in	11:05 am				
Volunteer Leads Candidates to Assigned Stations	11:35 am				
Enter Station	11:40 am	A	B	C	D
Exit Station	11:55 am				
Enter Next Station	12 pm	B	C	D	A
Exit Station	12:15 pm				
Enter Next Station	12:20 pm	C	D	A	B
Exit Station	12:35 pm				
Enter Next Station	12:40 pm	D	A	B	C
Exit Station Volunteer Leads Candidates to room to complete exit survey	12:55 pm				

Program Q & A – Round 1, 2 & 3 Candidates

1:20 pm – 1:50 pm with 2 interviewers
Candidates then complete Program Q & A exit survey

Round 4 – 2 pm – 3:15 pm

Event	Timing	Candidate 13	Candidate 14	Candidate 15	Candidate 16
Candidate Check-in	1:30 pm				
Volunteer Leads Candidates to Assigned Stations	1:50 pm				
Enter Station	2 pm	A	B	C	D
Exit Station	2:15 pm				
Enter Next Station	2:20 pm	B	C	D	A
Exit Station	2:35 pm				
Enter Next Station	2:40 pm	C	D	A	B
Exit Station	2:55 pm				
Enter Next Station	3 pm	D	A	B	C
Exit Station Volunteer Leads Candidates to room to complete exit survey	3:15 pm				

Round 5 – 3:20 pm – 4:35 pm

Event	Timing	Candidate 17	Candidate 18	Candidate 19	Candidate 20
Candidate Check-in	2:50 pm				
Volunteer Leads Candidates to Assigned Stations	3:15 pm				
Enter Station	3:20 pm	A	B	C	D
Exit Station	3:35 pm				
Enter Next Station	3:40 pm	B	C	D	A
Exit Station	3:55 pm				
Enter Next Station	4 pm	C	D	A	B
Exit Station	4:15 pm				
Enter Next Station	4:20 pm	D	A	B	C
Exit Station Volunteer Leads Candidates to room to complete exit survey	4:35 pm				

Program Q & A – Round 4 & 5 Candidates

4:50 pm – 5:20 pm 2 interviewers (opposite from lunch time)
Candidates then complete Program Q & A exit survey

APPENDIX D: Stakeholder Survey Candidate Consent Form

Child Life Studies Program
Department of Pediatrics
McMaster University
humphrc@mcmaster.ca

Quality Assurance Survey Consent Form - Candidates

You are invited to take part in a research survey about the multiple-mini interview process utilized in the Child Life Studies Program's candidate selection. We are interested in understanding candidate's perceptions of this interview format. This survey will take less than 10 minutes to complete.

There are no known risks associated with this survey. Taking part in this study is completely voluntary and will not have any bearing on this program's admissions selections. You are free to decline to participate, without consequence, or withdraw at any time prior to or at any point during the activity. Your responses will be anonymous and kept in a secured environment. Any report of this research that is made available to the public will not include your name as a participant or any other individual information.

Statement of Consent: I have read the above information, and have received answers to any questions. I consent to take part in the research study.

Participant name (please print): _____

Signature: _____

Date: _____

APPENDIX E: Applicant Survey

Applicant Questionnaire

MMI refers to multiple-mini interviews

1. **The interview process was well organized.**

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

2. **I had sufficient time to present my ideas in the MMI.**

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

3. **I could accurately portray myself during the MMI.**

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

4. **I personally prefer meeting 1 on 1 with an interviewer in each station instead of with a traditional interview panel.**

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

5. **Compared to traditional panel interviews, I found that the MMI was more fair.**

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

6. I enjoyed participating in the MMI.

1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

7. I am satisfied with the MMI as one of the tools for the admissions selection process into the Child Life Studies Program.

1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

8. Did you find any limitations with MMI?

Adapted from Oxford Grice (2014), Humphrey, et al. 2008, Razack, et al., 2009

APPENDIX F: Post - Program Question & Answer Period Survey

Post- Program Question & Answer Period Questionnaire

1. The program question and answer period was informative.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

2. My questions were addressed in this format.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

3. I have greater knowledge of the Child Life Studies Program and faculty.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

4. Were there any limitations to this question and answer period?

APPENDIX G: Stakeholder Survey Interviewer Consent Form

Child Life Studies Program
Department of Pediatrics
McMaster University
humphrc@mcmaster.ca

Quality Assurance Survey Consent Form - Interviewers

You are invited to take part in a research survey about the multiple-mini interview process utilized in the Child Life Studies Program's candidate selection. We are interested in understanding stakeholder perceptions of this interview format. This survey will take less than 10 minutes to complete.

There are no known risks associated with this survey. Taking part in this study is completely voluntary. You are free to decline to participate, without consequence, or withdraw at any time prior to or at any point during the activity. Your responses will be anonymous and kept in a secured environment. Any report of this research that is made available to the public will not include your name as a participant or any other individual information.

Statement of Consent: I have read the above information, and have received answers to any questions. I consent to take part in the research study.

Participant name (please print): _____

Signature: _____

Date: _____

APPENDIX H: Interviewer Survey

Interviewer Questionnaire

1. Interview stations are of the right duration.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

2. **We are measuring the right skills and abilities**.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

3. I had time to score applicants before the next applicant came into the room.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

4. **The scoring sheet allowed me to differentiate among applicants**.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

5. I would consistently offer the same assessment if asked to review my opinion at a later date.

1	2	3	4	5	6	7
Strongly disagree	Disagree	Disagree Somewhat	Neither Agree nor Disagree	Agree somewhat	Agree	Strongly Agree

6. **My interviewer workload is less in MMIs than traditional interviews**.

1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

7. **I personally prefer meeting 1 on 1 with an applicant in each station instead of being on a traditional interview panel.**

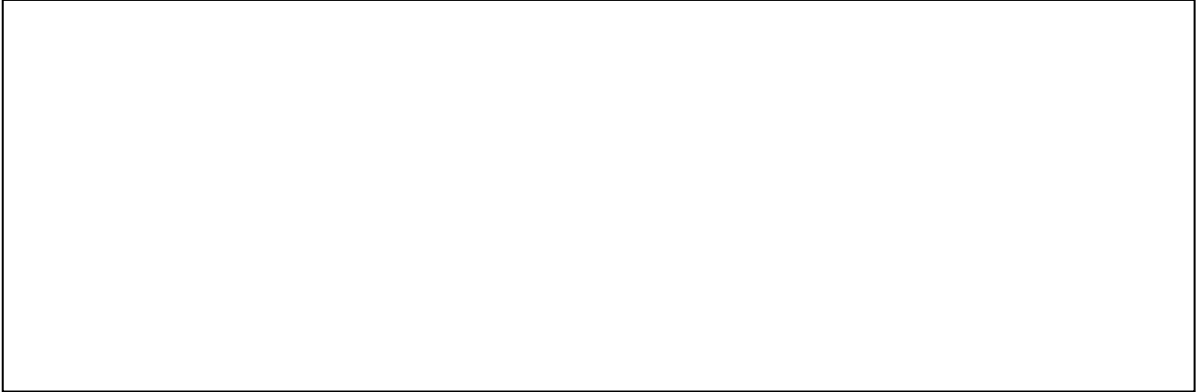| 1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

8. **Compared to traditional panel interviews, I found that the MMI was more**

   fair.

1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

9. **I would be in favour of continuing to utilize the MMI instead of traditional interviews.**

1 Strongly disagree	2 Disagree	3 Disagree Somewhat	4 Neither Agree nor Disagree	5 Agree somewhat	6 Agree	7 Strongly Agree

10. Did you find any limitations to MMI?

Adapted from Humphrey et al., 2008, Razack et al, 2009, Dore et al., 2010

References

Albanese, M.A, Snow, M.H., Skochelak, S.E., Huggett, K.N., Farrell, P.M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, 78(3), 313–321.

Axelson, R.D., Kreiter, C.D. (2009). Rater and occasion impacts on expected preadmission interview reliability. *Medical Education*, 43(12), 1198–1202.

Axelson, R., Kreiter, C., Ferguson, K., Solow, C., Huebner, K. (2010). Medical School Preadmission Interviews: Are Structured Interviews More Reliable than Unstructured Interviews? *Teaching and Learning in Medicine*, 22(4), 241-245.

Bloch, R., Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No.68. *Medical Teacher*, 34(11), 960-992.

Callwood, A., Cooke, D., Allan, H. (2014). Developing and piloting the multiple mini-interview in pre-registration student midwife selection in a UK setting. *Nurse Education Today*, 34(12), 1450-1454.

Cohen-Schotanus, J., Muijtjens, A.M.M., Reinders, J.J., Agsteribbe, J., van Rossum, H.J.M., & van der Vleuten, C.P.M. (2006). The predictive validity of grade point average scores in a partial lottery medical school admission system. *Medical Education*, 40(10), 1012–1019.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: A key to unlock professional assessment. *Medical Education, 36*(10), 972-978.

DeLisa, J. A., Jain, S. S., & Campagnolo, D. I. (1994). Factors used by physical medicine and rehabilitation residency training directors to select their residents. *American Journal of Physical Medicine and Rehabilitation*, 73(3), 152–156.

Dirschl, D. R., & Adams, G. L. (2000). Reliability in evaluating letters of recommendation. *Academic Medicine*, 75(10), 1029.

Dodson, M., Crotty, B., Prideaux, D., Carne, R., Ward, A., de Leeuw, E. (2009).The multiple mini-interview: how long is long enough? *Medical Education*, 43, 168–174.

Dore, K.L, Hanson, M., Reiter, H.I., Blanchard, M., Deeth, K., Eva, K.W. (2006). Medical school admissions: enhancing the reliability and validity of an autobiographical screening tool. *Academic Medicine*, 81 (10) (Suppl):70–73.

Dore, K.L., Kreuger, S., Ladhani, M., Rolfson, D., Kurtz, D., Kulasegaram, K, Cullimore, A.J., Norman, G.R., Eva, K.W., Bates, S., Reiter, H.I. (2010). The reliability and acceptability of the multiple mini-interview as a selection instrument for postgraduate admissions. *Academic Medicine: Journal of the Association of American Medical Colleges*, 85(10 Suppl), S60–S63.

Dotson Collins, D. (2013). Measuring Noncognitive Behaviors and Characteristics in Health Program Admissions Processes; Dissertation, University of Georgia.1-198.

Downing, S.M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, 38, 1006-1012.

Dunleavy, D.M., Whittaker, K.M. (2011). The evolving medical school admissions interview. *Association of American Medical Colleges Analysis in Brief,* 11(7), 1-2. https://www.aamc. org/download/261110/data/aibvol11_no7. pdf. Accessed November 20, 2015.

Edwards, J.C., Johnson, E.K., Molidor, J.B. (1990). The interview in the admission process. *Academic Medicine,* 65,167-177.

Elam C.L, Andrykowski, M.A. (1991). Admission interview ratings: Relationship to applicant academic and demographic variables and interviewer characteristics. *Academic Medicine,* 66(9), S13-15.

ETS Personal Potential Index (2016). Retrieved from https://www.ets.org/ppi. Accessed January 26, 2016.

Eva, K.W. (2003). On the generality of specificity. *Medical Education,* 37, 587-588.

Eva, K., Macala, C. (2014). Multiple mini-interview test characteristics: 'tis better to ask candidates to recall than to imagine. *Medical Education*, 48, 604–613.

Eva, K.W., Reiter, H.I. (2004b). Where judgment fails: pitfalls in the selection process for medical personnel. *Advances in Health Sciences Education*, 9, 161–174.

Eva, K.W., Reiter, H.,I., Rosenfeld, J., Norman, G.R. (2004c). The ability of the multiple mini-interview to predict pre-clerkship performance in medical school. *Academic Medicine*,79(10 suppl), S40–S42.

Eva, K.W., Reiter, H.I., Rosenfeld, J., Trinh, K., Wood, T.J., Norman, G.R. (2012). Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. J*AMA*, 308(21), 2233-2240.

Eva, K.W., Reiter, H.I., Trinh, K., Wasi, P., Rosenfeld, J., Norman, GR (2009). Predictive validity of the multiple mini-interview for selecting medical trainees. *Medical Education*, 43, 767-775.

Eva, K.W., Rosenfeld, J., Reiter, H.I., & Norman, G.R. (2004a). An admissions OSCE: the multiple mini-interview. *Medical Education,* 38, 314-326

Fraga, J.D., Oluwasanjo, A., Wasser, T., Donato, A., Alweis, R. (2013). Reliability and acceptability of a five-station multiple mini-interview model for residency program recruitment. *Journal of Community Hospital Internal Medicine Perspectives*, 3, 1-4.

Gabard, D.L, Porzio, R., Oxford, T., Braun, R. (1997). Admission Interviews: Questions of Utility and Cost in Masters of Physical Therapy Programs in the United States. *Physiotherapy Research International,* 2(3), 135–166.

Gafni, N., Moshinsky, A., Eisenberg, O., Zeigler, D., Ziv, A. (2012). Reliability estimates: behavioural stations and questionnaires in medical school admissions. *Medical Education*, 46(3), 277-288.

Gardner, K. (2014). Developing a Customized Multiple Interview for Dental Admissions. *Journal of Dental Education,* 78(4), 597-604.

Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: An exploratory meta-analysis. *Medical Teacher, 28*(4), 335-340.

Hanson, M.D., Kulasegaram, K.M., Woods, N.N., Fechtig, L., Anderson, G. (2012). Modified Personal Interviews: Resurrecting Reliable Personal Interviews for Admissions? *Academic Medicine*, 87(10), 1330-1334.

Harasym, P.H., Woloschuk, W., Mandin, H., Brundin-Mather R. (1996). Reliability and validity of interviewers' judgments of medical school candidates. *Academic Medicine,* 71, S40-S42.

Harris, S., Owen., C. (2007). Discerning quality: Using the multiple mini interview in student selection for the Australian National University Medical School. *Medical Education,* 41, 234-241.

Hecker, K., Donnon, T., Fuentealba, C., Hall, D., Illanes, O., Morck, D., Muelling, C. (2009). Assessment of Applicants to the Veterinary Curriculum Using a Multiple Mini-Interview Method. *Journal of Veterinary Medical Education*, 36(2), 166-173.

Hecker, K., Violato, C. (2011). A generalizability analysis of a veterinary school Multiple Mini Interview: effect of number of interviewers, type of interviewers, and number of stations. *Teaching and Learning in Medicine*, 23(4), 331-336.

Hofmeister, M., Lockyer, J., Crutcher, R. (2008). The acceptability of the multiple mini interview for resident selection. *Family Medicine,* 40, 734-740.

Hofmeister, M., Lockyer, J., Crutcher, R. (2009). The multiple mini-interview for selection of international medical graduates into family medicine residency education. *Medical Education*, 43(6), 573-579.

Humphrey, S., Simon, D., Wall, D., Diwakar, V., Goodyear, H.M. (2008). Multiple mini-interviews: Opinions of candidates and interviewers. *Medical Education,* 42, 207-13.

Husbands, A., Dowell, J.D. (2013). Predictive validity of the Dundee multiple mini-interview. *Medical Education,* 47, 717-725.

Jerant, A., Griffin, E., Rainwater, J., Henderson, M., Sousa, F, Bertakis, K.D., Fenton, J.J., Franks, P. (2012). Does Applicant Personality Influence Multiple Mini-

Interview Performance and Medical School Acceptance Offers? *Academic Medicine,* 87(9), 1250-1259.

Johnson, E.K., Edwards, J.C. (1991). Current practices in admission interviews at US medical schools, *Academic Medicine*, 66, 408–412.

Jones, P.E., Forister, J.G. (2011). A comparison of behavioural and multiple mini-interview formats in physician assistant program admissions. *The Journal of Physician Assistant Education*, 22(1), 36-40.

Jones, P.E., Simpkins, S., Hocking, J.A.(2014). Imperfect physician assistant and physical therapist admissions processes in the United States. *Journal of Educational Evaluation for Health Professions*, 11, 11. doi: 10.3352/jeehp.2014.11.11.

Joyner, P. U., Cox, W. C., White-Harris, C., & Blalock, S. J. (2007). The structured interview and interviewer training in the admissions process. *American Journal of Pharmaceutical Education, 71*(5), 1-5.

Kirchner, G.L., Holm, M.B. (1997). Prediction of academic and clinical performance of occupational therapy students in an entry-level master's program. *The American Journal of Occupational Therapy*, 51, 775-779.

Klehe, U-C, Latham, G. (2006). What would you do - really or ideally? Constructs underlying the behaviour description interview and the situation interview in

predicting typical versus maximum performance. *Human Performance,* 19(4), 357-382.

Knorr, M., Hissbach, J. (2014). Multiple mini-interviews: same concept, different approaches. *Medical Education,* 48, 1157–1175.

Kreiter, C. D., & Kreiter, Y. (2007). A validity generalization perspective on the ability of undergraduate GPA and the medical college admission test to predict important outcomes. *Teaching and Learning in Medicine,* 19(2), 95–100.

Kreiter, C.D., Axelson, R.D. (2013). A perspective on medical school admission research and practice over the last 25 years. *Teaching and Learning in Medicine*, 25 Suppl 1, S50-6.

Kreiter, C.D., Yin, P., Solow, C., Brennan, R.L. (2004). Investigating the Reliability of the Medical School Admissions Interview. *Advances in Health Sciences Education*, 9, 147-159.

Kulatunga-Moruzi C., Norman, G.R. (2002). Validity of admission measures in predicting performance outcomes: The contribution of cognitive and noncognitive dimensions. *Teaching and Learning in Medicine*, 14, 34–42.

Kumar, K., Roberts, C., Rothnie, I., du Fresne, C., Walton, M. (2009). Experiences of the multiple mini-interview: a qualitative analysis. *Medical Education*, 43, 360-367.

Kyllonen, Patrick (n.d.). The Research Behind the ETS ® Personal Potential Index (PPI) Retrieved from https://www.ets.org/Media/Products/PPI/10411_PPI_bkgrd_report_RD4.pdf

Lemay, J.F., Lockyer, J.M., Colin, V.T., Brownell, A.K. (2007). Assessment of non-cognitive traits through the admissions multiple mini interview. *Medical Education*, 41, 573-579.

Meredith, K.E., Dunlap, M.R., Baker, H.H. (1982). Subjective and objective admissions factors as predictors of clinical clerkship performance, *Journal of Medical Education*, 57, 743–751.

Norman, G., Evan, K. (2010). Quantitative research methods in medical education. In Swanick, T. (ed.). *Understanding Medical Education: Evidence, Theory & Practice*. (p. 301-322). Blackwell Publishing: Oxford.

O'Brien, A., Harvey, J., Shannon, M., Lewis, K., Valencia, O. (2011). A comparison of multiple-mini interviews and structured interviews in a UK setting. *Medical Teacher*, 33(5), 397-402.

Oxford Grice, K. (2014). Use of multiple mini-interviews for occupational therapy admissions. *Journal of Allied Health,* 43(1), 57-61.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education*, 46, 850–868.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., Cleland, J. (2016). How effective are selection methods in medical education? A systematic review. *Medical Education*, 50(1), 36-60.

Patterson, F., Lievens, F., Kerrin, M., Zibarras, L., Carette, B. (2012). Designing Selection systems for Medicine: The importance of balancing predictive and political validity in high-stakes selection contexts. *International Journal of Selection and Assessment*, 20(4), 486-496.

Pau, A., Jeevaratnam, K., Chen, Y.S., Fall, A.A., Khoo, C. Nadarajah, V.D. (2013). The Multiple Mini-Interview (MMI) for student selection in health professions training – A systematic review. *Medical Teacher*, 35, 1027-1041.

Perkins, A., Burton, L., Dray, B., Elcock, K. (2013). Evaluation of a multiple-mini-interview protocol used as a selection tool for entry to an undergraduate nursing programme. *Nurse Education Today,* 33(5), 465-469.

Phillips, A.W., Garmel, G.M. (2014). Does the Multiple Mini-Interview Address Stakeholder Needs? An Applicant's Perspective. *Annals of Emergency Medicine*, 64(3), 316-319.

Price, P.B, Lewis, E.G., Loughmiller, G.C, Nelson, D.E., Murray, S.L., Taylor, C.W.
(1971). Attributes of a good practising physician. *Journal of Medical Education*,
46(3), 229–37.

Prideaux, D., Roberts, C., Eva, K., Centeno, A., McCrorie, P., McManus, C., Patterson,
F., Powis, D., Tekian, A., Wilkinson, D. (2011). Assessment for selection for the
health care professions and specialty training: Consensus statement and
recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3),
215-223

Raghavan, M., Burnett, M, Martin, B.D., Christensen, H., Young, D.G., Mackalski, B.,
Aoki, F. (2013). Utility of a writing station in the multiple mini-interview.
Journal of Veterinary Medical Education, 40 (2), 177–83.

Razack, S., Faremo, S., Drolet, F., Snell, L., Wiseman, J., & Pickering, J. (2009).
Multiple mini-interviews versus traditional interviews: stakeholder acceptability
comparison. *Medical Education*, 43, 993–1000.

Reiter, H., Eva, K., Rosenfeld, J., Norman, G. (2007). Multiple mini-interviews predict
clerkship and licensing examination performance. *Medical Education,* 41, 378–
384.

Reiter, H.I. & Eva, K.W. (2005). Reflecting the relative values of community, faculty,
and students in the admissions tools of medical school. *Teaching and Learning
in Medicine*, 17(1), 4–8.

Reiter, H.I., Salvatori, P., Rosenfeld, J., Trinh, K., Eva, K.W. (2006). The effect of defined violations of test security on admissions outcomes using multiple mini-interviews. *Medical Education*, 40, 36-42.

Roberts C, Rothnie I, Zoanetti N, Crossley J. (2010) Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education*, 44, 690–8.

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., Tiller, D. (2008). Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical Education*, 42, 396–404.

Roberts, C., Zoanetti, N., Rothnie, I. (2009). Validating a multiple mini-interview question bank assessing entry level reasoning skills in candidates for graduate-entry medicine and dentistry programmes. *Medical Education*, 43, 350–9.

Rosenfeld, J.M., Reiter, H.I., Trinh, K., Eva, K.W. (2008). A cost efficiency comparison between the multiple mini-interview and traditional admissions interviews. *Advances in Health Sciences Education: Theory and Practice*, 13(1), 43-58.

Salvatori, P. (2001). Reliability and Validity of Admissions Tools Used to Select Students for the Health Professions. *Advances in Health Science* Education, 6, 159–175.

Schuwirth, L.W., & van der Vleuten, C.P.M. (2011). General overview of the theories used in assessment: AMEE guide no. 57. *Medical Teacher, 33*(10), 783-797.

Sebok, S.S., Luu, K., Klinger, D.A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. *Advances in Health Sciences Education*,19, 71-84.

Serres, M.L., Peeters, M.J. (2012). Overcoming content specificity in admission interviews: the next generation? *American Journal of Pharmaceutical Education*, 76 (10), 207.

Siu, E., Reiter, H. (2009). Overview: what's worked and what hasn't as a guide towards predictive admissions tool development. *Advances in Health Sciences Education*, 14, 759-775.

Standridge, J., Boggs, K.J., & Mugan, K.A. (1997). Evaluation of the health occupations aptitude examination. *Respiratory Care*, 42 868-872.

Stansfield, R.B., Kreiter, C.D. (2007). Conditional reliability of admissions interview ratings: extreme ratings are the most informative. *Medical Educator*, 41, 32–38.

Streiner, D.L. Norman, G.R., Cairney, J. (2015). *Health Measurement Scales: A practical guide to their development and use.* Fifth edition. Oxford University Press: Oxford, UK.

The Association for the Study of Medical Education. (2014). In Swanick, T. (ed.) *Understanding Medical Education: Evidence,Theory and Practice*. Second edition. John Wiley & Sons Ltd.:West Sussex, UK

Tiller, D., O'Mara, D., Rothnie, I., Dunn, S., Lee, L., Roberts C. (2013). Internet-based multiple mini-interviews for candidate selection for graduate entry programmes. *Medical Education*, 47, 801–10.

Uijtdehaage, S., Doyle, L., Parker, N. (2011). Enhancing the Reliability of the Multiple Mini-Interview for Selecting Prospective Health Care Leaders. *Academic Medicine*, 86 (8) 1032-1039.