**Why are you interested in the particular area of research?**

My interest for non-cognitive, or to be more specific, personality assessments, was not built in a day. ~~Before I first read papers in graduate school on how personality is one the most important predictors of educational performance, and job performance, I had begun to realize, although intuitively, that to succeed, we need more than a high IQ, or a high math score.~~

**It is a great predictor: as long as there’s a job that needs people for it, there’s a need for good personality tests.**

**I realize this both intuitively and empirically.**

**Intuitively:**

I got my bachelor’s degree from Tsinghua University, which is one of the two top schools in China, the other being Beijing University. To get into Tsinghua, you need to be the top 1% in your province on the very hard College Entrance Examinations, and that is to say, Tsinghua is filled up with the smartest people in the country, but that’s only the case for undergraduate school. In graduate school at Tsinghua, there are a large number of students graduating from a variety of universities across China, some of which are by no means top-notch, yet many of these students I know have accomplished great achievement in academia. Whereas about every year, some super smart undergraduate students in Tsinghua get dismissed for failing to many courses, one of whom used to be my roommate. When I looked at these two groups of people 6 years ago, I realized that those who came from an ave6rage undergraduate school but managed to thrive at Tsinghua were with no exception hard working, curious about knowledge, and never tired of exploring. Those undergraduates who got dismissed, including my ex-roommate, however, were almost all because of being too addicted to computer games to attend classes and fulfill course requirements. This is not about their intelligence, I thought to myself, this about their… Style. Thanks to Goldberg, I later learnt that this “style” is actually part of our personality, and thanks to the many papers I read in graduate school, this “style”, or personality, is one of the most important and consistent predictor of both educational and job performance.

**Empirically:**

In a project that I just finished with Professor Brent Roberts last year, the PISA 2012 data was used to examine across 9 major cultural groups consisting of all 68 participating countries and regions in PISA 2012, the prediction of math achievement and absenteeism from school by perseverance, controlling for SES and gender. Perseverance is a facet of conscientiousness, which is one of the most stable and strongest predictor of academic performance among the Big-Five personality traits. The analysis was carried out in Mplus using multigroup SEM. It was found that perseverance, especially the negative factor (including items such as giving up easily, and putting off difficult problems) of perseverance, was a great predictor of low math achievement and high truancy, and this relationship held across all 9 cultural groups. I witnessed, through this project with global large-scale data, how personality can actually influence important aspects of the academic performance of high school students all over the world, regardless of cultural background, SES, and gender.

**Through reading:**

Personality assessments are not only tools of good criterion validity, but also tools of building a fair hiring process offering equal employment opportunity for everyone.

After I got into the I-O program at UIUC, I learnt about adverse impact, and started to understand the contribution of personality assessments to reducing adverse impact in the hiring process, especially in the United States, a country due to complex social and historical reasons, consisting of people of different ethnicity coming from drastically different socio-economic groups.

**Conclusion:**

A good personality assessment is a combination of high criterion validity (accuracy) and low adverse impact (fairness), and is essential to personnel selection, which is one of the key aspects to a successful organization. However, to build good personality assessments that are up to date, there are issues that need to be settled. Some of these topics I’ve had research experience with, some I haven’t, but I am extremely passionate about expanding my knowledge about all these issues, as well as exploring potential solutions for them.

**The modeling of personality data ~~and differential item functioning (DIF)~~**

Like studies carried out within a single group, to examine DIF with data from multiple groups, it should first be decided what model to use. There has long been a heated debate over when an IRT approach is adopted, whether a dominance models (e.g., 2PL, SGR) or an ideal point model (e.g., GGUM) is more appropriate for studying personality data. Many studies have been published focusing on single-group personality data, and few attempts have been made comparing the two types of models in a DIF analysis.

Therefore, my Master’s thesis adopted an IRT approach to studying DIF of the Well-being scale and the Curiosity scale from the Comprehensive Personality Scale (CPS), using data collected from China and the United States. A dominance IRT model (i.e., SGR) and an ideal point model (i.e., GGUM) were applied to the data, with the log-likelihood ratio (LR) as the test statistic for model comparison and DIF detection. A DIF effect size measure was also applied in order to obtain the magnitude of DIF, and thus compensate for the oversensitivity of the **Null Hypothesis Significance Testing** (**NHST**) to large samples. It was found that GGUM in general had moderately better model-data fit than SGR, due to the fact that SGR is terrible at modeling items of medium extremity, and that GGUM is able to handle items of all levels of extremity. Test differential functioning (TDF) obtained by combining the test characteristic curves (TCCs) of the two groups for each of the scales showed that both scales exhibited smaller DTF under GGUM than SGR, which means according to GGUM, both the Well-being and the Curiosity scales are more equivalent cross cultures than when they are examined with SGR. DIF couldn’t be examined for GGUM due to ill-conditioned matrices produced by GGUM2004 using the marginal maximum likelihood (MML) technique, but the results on TDF displayed GGUM’s potential to be the model to use in DIF analyses for personality tests, considering that in modeling personality data, GGUM is more flexible, and does an even better job than SGR.

The three questions I have after this project, and hope to find answers or get inspired at ETS are: first, why some of the items that were deemed intermediate by experts did not turn out to have a bell-shaped curve? Two, apart from MML, are there alternative estimation techniques for the GGUM that will give more accurate results, so that DIF could be computed? I know about the Bayesian estimation, but are there other choices?

**The forced-choice personality tests**

I’m also very intrigued by the forced-choice response format and its application to personality tests. Compared with a Likert scale, tests adopting this innovative response format will see fewer biases induced by faking, especially in a high-stake situation, and by different response styles of participants due to their cultural backgrounds, both of which are very commonly seen criticism of Likert scales. It’s not hard to imagine that multidimensional forced-choice (MFC) personality tests will help organizations make wiser selection decisions by providing more accurate test scores, well, as long as there’s a good way of scoring those tests.

Stark et al. (2005) presented the multidimensional forced-choice pairwise preference (MUPP) model based on the GGUM for scoring, and is adopted in the WorkFORCE™ Assessments, while Brown and Maydeu-Olivares (2011) introduced the Thurstonian IRT model that intended only for dominance items. The research on forced-choice response models are still at the very beginning, and there are lot to explore in this field. Besides the various models, for example, there are researchers such as Montangas, who are trying to bring ipsative scores back to life for forced-choice tests (FCT). Although it has been widely believed that the traditional scoring method of adding up the reversal of the rank orders of all items is a dead end for FCTs, as these ipsative scores are making it impossible to compare between individuals, recently, some researchers are giving it another though. Hontangas and colleagues (2015; 2016), by using the MUPP model and extending it to more than two items per block and different forced-choice (FC) formats, have found through two simulation studies, that under both the ideal point model or the dominance model, the traditional ipsative scores can actually have acceptable correlations with the true scores under certain circumstances, given that the type of the forced-choice format is RANK (i.e., all the items are ranked in the order of their descriptiveness of the the participants). Not like scoring, where at least some models and methods have been proposed, measurement equivalence studies are no where to be found for forced-choice tests. I haven’t had a lot of experience working with FCTs, but I do have a strong feeling that the future of personality tests lies in this young research area with a lot of mysteries and possibilities, and I just can’t wait to be part of it.

And through an internship at ETS, I’ll have the opportunity to be part of the projects that target these important issues of personality assessments, from an academic perspective, as well as a practical one. I’ll have the

**Why you want to come to ETS?**

I want to come to ETS because I believe it will help me to achieve my goal to add to my knowledge and experience in the application of item response theory in noncognitive assessments. ETS has the

ETS is one of the first to invent item response theory, which is guiding and underlying my research.

I want to come to ETS because of its long-term world-wide reputation in psychometrics, and its continuous effort in the field. It’s an organization with the wisdom of a 70-year-old and the curiosity of a 7-year-old. As one of the world’s largest testing and assessment company, ETS is never satisfied with what it has achieved: from the paper-based GRE test to the computerized question-adaptive test until Aug 2011, to today’s multistage section-adaptive test, ETS is constantly exploring paths to a more accurate, secure, and user-friendly test. In the area of noncognitive assessments, ETS has has impressed me with its rigorous efforts in applying the GGUM and the forced-choice format, which is still at its infancy, to the development and validation of the innovative WorkFORCE assessments. This spirit of exploration is what I value a lot about ETS, where research is the fuel of its products, and the products in turn examine the quality of the research. Therefore, I believe ETS is the best place for me to achieve my goals of

adding to my knowledge and research experience in psychometrics, as well as applying to the real world what I have learnt in the lab.

**What skills and knowledge will you bring to the internship?**

**Item response theory and programming**

I have worked closely with some of the most widely used IRT models (including 2PL, SGR, the GGUM) to study response data of personality tests. I have proficiency in using a variety of IRT-related software including BILOG, MULTILOG, GGUM2004, MODFIT, flexMIRT, and MCMC GGUM for item calibration, test scoring, model-data fit evaluation, and DIF analyses.

I started working with R in undergraduate school, and have proficiency of using R for simulation programming, and automated programming for the constrained and free baseline approaches to DIF analyses.

**Public speaking**

I don’t fear speaking in front of a crowd. When I give a presentation, my goals are delivering clearly the information I want to deliver and trying my best to make sure that everyone has a good time. I did a good job on my first-year presentation, where I, as a first-year PhD student, presented my meta-analysis on narcissism in front of more than 40 professors and more senior PhD students in the department. I’ve been teaching undergraduate courses ever since I came to UIUC in 2013. I’ve taught both labs of 20 students and lectures of 100, so I’m pretty comfortable with public speaking. When I’m giving a talk, I hold an honest attitude, so that the possibility of being judged or of not knowing the answer to a question from the audience never bothers me.

**Understanding of different cultures**

I’ve taken two of Professor Dov Cohen’s seminars on cultural and social psychology, and have a broad understanding of how culture influences almost every aspect of our lives, some of the aspects that relate to my research interest in cross-cultural personality assessments are self-regards (i.e., how people view themselves), comparison, socially-desired values, and so on.

I believe my knowledge of culture will likely to help me with understanding and explaining results obtained from cross-cultural personality assessments, and it will also give me more confidence working in the diverse environment at ETS.

**Communication skills**

Verbal, written, and interpersonal.

**Verbal: Proficient at delivering oral presentations**

Public speaking

I don’t fear speaking in front of a crowd. When I give a presentation, my goals are delivering clearly the information I want to deliver and trying my best to make sure that everyone has a good time. I did a good job on my first-year presentation, where I, as a first-year PhD student, presented my meta-analysis on narcissism in front of more than 40 professors and more senior PhD students in the department. I’ve been teaching undergraduate courses ever since I came to UIUC in 2013. I’ve taught both labs of 20 students and lectures of 100, so I’m pretty comfortable with public speaking. When I’m giving a talk, I hold an honest attitude, so that the possibility of being judged or of not knowing the answer to a question from the audience never bothers me.

Written: Skilled in conveying information in a professional and timely manner

On average, I deal with more than 10 emails a day, including emails from the department, advisors, projects collaborators, and students. The content of the email covers school affairs, academic research, course requirements, grades and so on. I’ve had a lot training from these in this area, and a Blackberry, allows me to respond to emails quickly.

Interpersonal:

I’ve held brainstorm and discussion sessions for item writing for a personality assessment development project, where I was responsible for motivating discussions, and steering the conversations back to the topic from time to time. The experience of running discussion sessions, joining regular seminars and journal club, as well as everyday interactions with my undergraduate students has strengthened various interpersonal skills of mine.

For example, when running discussion or participating seminars, I need to listen carefully others’ opinions, ask engaging and relevant questions, and remember to respect others by not interrupting them, especially when the conversation gets heated. When I’m talking with my students, however, my major task will be first listening to them, and then craft thoughtful and informative responses, which are what they were looking for from me.

**What contributions do you believe you can make to a project in that research area?**

**My research experience equipped me with the skills necessary to a project:**

First of all, I have research experience relevant to noncognitive assessments, especially in an item response theory framework. My Master’s thesis adopted an IRT approach to studying DIF of the Well-being scale and the Curiosity scale from the Comprehensive Personality Scale (CPS), using data collected from China and the United States. A dominance IRT model (i.e., SGR) and an ideal point model (i.e., GGUM) were applied to the data, through both the constrained and the free baseline approaches, with the log-likelihood ratio (LR) as the test statistic for model comparison and DIF detection. A DIF effect size measure was also applied in order to obtain the magnitude of DIF, and thus compensate for the oversensitivity of the **Null Hypothesis Significance Testing** (**NHST**) to large samples. I used MULTILOG and GGUM2004 for modeling the data. Such experience has not only added to my understanding of item response theory, but has also greatly improved my skills of running a variety of software necessary for IRT analyses. I am proficient in R programing for IRT analyses, including running packages for DIF analyses (e.g., “lordif”, “difR”), data generation based on different IRT models, writing functions that automatically generate syntax for IRT software such as MULTILOG, and GGUM2004, and calling external IRT software to run an automated analysis such as the constrained baseline and the free baseline modeling. I am also capable of transforming MATLAB code to R code, as I did with the DIF effect size measure syntax. I am also very familiar with MODFIT, the Excel macro for computing the fit of IRT models, and I am able to quickly get IRT model fit results computed and give accurate interpretation of them for many major IRT models (e.g., 2PL, 3PL, SGR, GGUM…). I am very familiar with ICCs, the visual aids for item evaluation, as in the course of working on my thesis, I saw and interpreted over a hundred of them, for the purposes of detecting non-discriminant items, unfolding items, and negative items.

During a project that I just finished with Professor Brent Roberts last year, I had the opportunity of using CFA, a different approach than IRT to study measurement equivalence across cultures. The project used the PISA 2012 data to examine the prediction of math achievement and absenteeism from school by perseverance, controlling for SES and gender, across 9 major cultural groups consisting of all 68 participating countries and regions in PISA 2012. The analyses were conducted through a multigroup structural equation model (SEM), where the measurement model was a correlated factor model, with the two factors being negative and positive factors of perseverance, respectively. The analysis was first carried out in AMOS, and I turned to Mplus after finding out that AMOS couldn’t compute factor means. In PISA 2012, there is a wide variety of sample sizes, ranging from 293 for Liechtenstein to 33806 for Mexico, so in order to avoid potential biases brought by the drastically different sample sizes, we decided to randomly sample 500 people from each country/region for the analyses (i.e., Liechtenstein was dropped from the analysis). This was done in SPSS, where I performed a variety of basic analyses to prepare the data for SEM, including data split and merging, factor analyses, correlations by and across groups. Through this project, I gained hands-on experience with modeling measurement equivalence using large-scale data with AMOS, Mplus, and SPSS. I also was responsible for grouping the 68 countries into 9 major cultural groups, which was done partially based on Saucier and colleagues’ (2005) grouping of 38 countries, but I felt that their way of grouping, which considered only geographic proximity, was not enough for the form of groups that are representative and discriminate from each other in terms of culture, I took a step further, and consulted The United Nations Statistics Division (UNSD) website, as well as Wikipedia, and eventually came up with a grouping strategy that took into consideration not only proximity of the countries, but also other factors such as cultural background, political systems, and economy conditions. I’m not only good with numbers and programs, I’m also good with searching, organizing, and processing information that’s closely related to the society and the world.

In another ongoing project with Professor Brent Roberts, we were trying to come up with a new version of the Chernyshenko Conscientiousness Scale, with more facets covered, but fewer items contained in each facet. In order to develop a new scale, we needed to first create a relatively large item pool, so I would hold discussion session in the lab, where every member came up with new items for each of the facets, and talk about them and decide if they were clear and accurate enough to be kept in the item pool. Sometimes, I needed to be the one to initiate the discussion, and motivate everyone to participate, especially during meetings early in the morning, while sometimes I needed to steer discussion back to the topic. While the discussion went on, I was also the responsible for putting down the new items along with lab members’ comments on them. This experience equipped me with the communication skills necessary for a group project, and the ability of information selection and organization. Therefore, when there’s a new project, I will have no problem leading discussion or brainstorm sessions within the group, while putting down important information in an orderly fashion for later.

Apart from the skills I’ve acquired through my research experience, I do believe that some of the qualities of mine will contribute to a project in the research area.

One thing I’ve learnt from taking and teaching my social psychology classes is the danger of conformity. I’m a team player, but I don’t conform.

Ever since I got into graduate school, I’ve worked with a variety of more senior PhD students and faculties, both locally and remotely. I’ve had a great time collaborating with all of them and have learnt a great deal from them, but that doesn’t mean I’ll hesitate asking questions when something is not right. For example, during my first year, when I was coding in MULTILOG for a DIF study, I was not sure if I should put down 1 or 2 for the number of groups, because I had seen both. A senior student in the lab told me to put down 1 as that’s what he and some others he knew had been using. He put it as it was common sense. I was not persuaded at all by the reason that “everyone else is using 1”. After consulting with other more experienced researchers and looking it up in papers and research reports, I finally confirmed that the correct group number was supposed to be 2, instead of 1, although the code for equating item parameters implicitly identified two groups for MULTILOG. I privately informed the student in the lab about the correct group number, and fortunately, he hadn’t published any of those projects yet. What was interesting was that both sides I consulted thought their answer was the obviously correct one, and had never given it too much thoughts before I asked. I think what I can contribute to a project is that I don’t take any conclusions for granted, regardless of whether it’s from a colleague or a supervisor, and whether it’s a considered by many an obvious one. If I find something confusing, I’ll ask.

I’m also an explorer, and I like to share my findings. In the same DIF project mentioned above, when I was running the baseline models in R under GGUM, I needed to call an external GGUM2004 program from R. To start the program, I needed to use the “shell” command in R to first call a .au3 file written by a computer sciences student several years ago, which initiates the command file of GGUM2004. This .au3 file is handed down in the lab by former students, but the problem with it we’ve been having recently in the lab is that this “shell” command is working only on the computer of one student in our lab, and he had no idea why, and the technical group on campus was of much help, either. This student kindly offered his computer to me when I needed to run a simulation study with the “shell” command. As grateful as I was, I said no. This was because I wanted to solve this program problem once and for all, rather than avoiding it, and I was confident that I could solve this problem, because with a close check, I believe the problem lies in the syntax of the .au3 file used to call GGUM2004. Next, I spent 3 days learning about basic .au3 syntax, and finally located the source of the problem – the file path is wrong, and the cmd program is missing from the assigned folder. I also added two lines of code that enables the cmd window of GGUM2004 to close to let the simulation continue if there’s a singular matrix. Under the same circumstance, with the old .au3 code, the window won’t close itself and the automated process will be paused. I shared my findings and the code with the lab, and now everyone in our lab can easily get the “shell” command to run on any PC with GGUM2004 installed. I am perseverant in solving problems encountered in the course of research, and I believe this quality of mine will contribute to personality projects, especially those about exploring alternative scoring for the forced-choice responses, which is a relatively new area with a lot of unanswered questions and mysteries.