

SAMPLE PORTFOLIO: Data Scientist

BROAD OVERVIEW OF RESEARCH PROJECTS:

I played a lead role in the research design, survey sampling, data collection, and statistical analysis of numerous research projects. In addition to overseeing multiple research teams (up to 27 team members), I played a major role in the survey design, sampling selection, and data compilation of these projects. Additionally, I implemented procedures such as missing data imputation, weighting procedures, and bias reduction techniques to ensure data representativeness and accuracy. I also constructed complex predictive models with advanced statistical methods (e.g., structural equation modeling).

Postdoctoral Research Scientist, Columbia University:

- **Fragile Families**: large-scale longitudinal study on children born to low-income, single-parent families (n = 6,000 participants)
- **myPersonality**: study looking at the links between personality and Facebook networks and behavior (n = 7.5 million participants)
 - Python (pandas)

Graduate Researcher, University of Maryland:

- **The Friendship Project**: longitudinal study on youth's social relationships and adjustment from elementary to high school (n = 1000 participants)
- **Traumatic Brain Injury Study**: examines the social behavior and adjustment of children with Traumatic Brain Injury (n = 50 participants)
- **Multi-method Intervention for Socially Anxious Preschoolers**., examines the effectiveness of an intervention/prevention program for socially anxious preschoolers (n = 15 participants, data collection ongoing)
- **Bullying in Hong Kong Schools**: international study looking on school bullying from the 3rd grade through the 8th grades in Hong Kong schools (n = 1200 participants)
- **IEA Civic Engagement Study**: the largest and most rigorous international study of civic education. It looks at the links between citizenship, national identity, school characteristics, and social cohesion in youth across 28 countries (n = 90,000 participants).
- **Early Childhood Longitudinal Study (ECLS)**., longitudinal study conducted by the U.S. Department of Education on early education in the U.S. (n = 30,000 participants)
- **The Chinese Youth Study (Lead Principal Investigator)**: I was awarded a National Science Foundation fellowship to conduct an international study on the links between personality, social relationships, and adjustment in young Chinese adolescents living in Beijing, China (n = 300 participants)
- **Understanding Taiwanese Youth Study (Lead Principal Investigator)**. The study looks at the links between personality, social relationships, and loneliness in young adolescents living in Taipei, Taiwan (n = 250 participants)

SPECIFIC EXAMPLES (more details available upon request):

- **SURVEY DESIGN:**
 - **The Chinese Youth Study:** As the Lead Principal Investigator, I independently lead all aspects of project's survey design, including sample selection, survey and instrument design and development, data collection protocols, and statistical analysis.
 - I designed **10** survey questionnaires and measurement instruments to examine Chinese youth's personality and social relationships (full surveys available upon request)
 - In making survey design decisions, I reviewed existing literature on survey instruments, collaborated with survey methodologists and researchers, conducted numerous focus groups in the U.S. and China, and engaged in efforts to ensure measurement quality (e.g., rigorous procedures for translating and backtranslating Western-based measures into Chinese).
 - Additionally, I conducted psychometric studies using advanced statistical analytics (e.g., confirmatory factor analysis; IRT) to ensure the reliability and validity of survey instruments. These studies have since been published in peer-reviewed journals.
 - **Examples of deliverables:**
 - Peer-reviewed journal articles:
 - Wang, J. M. (2015). Preference-for-solitude and depressive symptoms in Chinese adolescents. *Personality and Individual Differences*. Advanced online publication. doi.org/10.1016/j.paid.2015.09.033
 - Wang, J. M. (2014). Psychometric properties of the Network Relationship Inventory- Social Provision Version in Chinese youth. *Child Psychiatry and Human Development*. Advanced online publication. doi:10.1007/s10578-014-0438-6.
 - **Understanding Taiwanese Youth Study:** As the Lead Principal Investigator, I independently lead all aspects of research design, including sample selection, survey and instrument design and development, data collection protocols, and statistical analysis.
 - I designed **11** survey questionnaires to examine Taiwanese youth's adjustment and well-being (full surveys available upon request).
 - In making survey design decisions, I reviewed psychometric findings from my previous Chinese Youth Study. I also conducted additional literature reviews on test measurements and survey instruments, collaborated with survey methodologists and researchers, and engaged in efforts to ensure measurement quality (e.g., rigorous procedures for translating and backtranslating Western-based measures into Chinese).

- Additionally, I conducted psychometric studies using advanced statistical analytics (e.g., confirmatory factor analysis; IRT) to ensure the reliability and validity of survey instruments and questionnaires. These studies have since been published in peer-reviewed journals.
- **Examples of deliverables:**
 - Peer-reviewed journal articles:
 - Marlies, M., Wang, J. M., & Goosens, L. (under review). The loneliness and aloneness scale for children and adolescents: Measurement invariance among Belgian and Taiwanese adolescents.
- **SURVEY SAMPLING:** To obtain accurate representation of populations, high response rates, and quality responses, all sampling procedures were conducted according to best practices guidelines for survey sampling (Dillman, 2000; Stapleton).
 - **Early Childhood Longitudinal Study (ECLS):** A total of 30,000 participants were surveyed. Participants were selected using **stratified sampling criteria**, with children from each ethnic group (White, Asian, Black, American Indian and Alaska Native, Native Hawaiian or Other Pacific Islander) and level of socioeconomic status (low, middle, high) selected from 200 public and private kindergartens
 - **Examples of deliverables:**
 - Wang, J. M. & Wimsatt, M. (under review). Young children's internalizing problems: Do self-control and parent-teacher meetings matter?
 - **Fragile Families:** A total of 6,000 participants were surveyed. All participants were selected using **stratified random sampling** of all U.S. cities with 200,000 or more people. Stratification decisions were made according to policy environments and labor market conditions in the different cities. The sampling occurred in three stages: 1) state-level 2) city-level 3) hospital-level.
 - **Examples of deliverables:**
 - Brief reports:
 - <http://www.fragilefamilies.princeton.edu/briefs2.asp>
 - Invited talk (available upon request)
 - **The Chinese Youth Study:** I wanted to select a sample from a city in China that is reflective of the rapid modernization and Westernization that China is experiencing. As the capital of the country, Beijing is an ideal choice given its urban location and its fast industrial growth.
 - In selecting the sample, I first identified the characteristics that I wanted the sample to reflect (e.g., urban environment, middle-class, predominantly Han Chinese). I then collaborated with school officials in Beijing to select schools that best fit these criteria.

- **Examples of deliverables:**
 - Peer-reviewed journal article:
 - Wang, J. M. (2015). Preference-for-solitude and depressive symptoms in Chinese adolescents. *Personality and Individual Differences*. Advanced online publication. doi.org/10.1016/j.paid.2015.09.033
 - Wang, J. M. (2014). Psychometric properties of the Network Relationship Inventory- Social Provision Version in Chinese youth. *Child Psychiatry and Human Development*. Advanced online publication. doi:10.1007/s10578-014-0438-6.
 - **Understanding Taiwanese Youth Study:** I wanted to select a sample that is reflective of the typical middle school environment in an urban setting in Taiwan. Taipei, given its central location and urban environment, was the ideal representation of this population.
 - In selecting the sample, I first identified the characteristics that I wanted the sample to reflect (e.g., urban environment, middle-class, predominantly Taiwanese). Then, I collaborated with school officials in Taipei to select schools that best fit these criteria.
 - **Examples of deliverables:**
 - Professional conference presentations:
 - Wang, J. M. (2014, July). Shyness and rejection sensitivity in Chinese youth: The role of aversion to aloneness. In B. Chen & Y. Xu (Chair), *Contextual Influences on the Development of Social Withdrawal*. Symposium conducted at the biennial International Society for the Study of Behavioural Development, Shanghai, China.
- **WEIGHTING:**
 - **Fragile Families:** A total of 6,000 participants were surveyed. Several different weighting procedures were used to ensure the representativeness and accuracy of our data. Weights were created and included in analyses at the **national level** to ensure that data from the 16 randomly selected cities were representative of the births that had occurred in large U.S. cities during those periods. Weights were also created and included in analyses at the **city-level** to ensure that data from all 20 cities were representative of births in these cities. Additionally, attrition weights were created using logistic response models to account for attrition.
 - **Examples of deliverables:**
 - Brief reports:
 - <http://www.fragilefamilies.princeton.edu/briefs2.asp>
 - Invited talk (available upon request)

- **VARIANCE ESTIMATION:**

- **Early Childhood Longitudinal Study (ECLS):** Given its complex research design, various methods were used to adjust standard errors and correctly estimate variance, including the **jackknife replication** variance estimation and the **Taylor series linearization method**.

- **Examples of deliverables:**

- Wang, J. M. & Wimsatt, M. (under review). Young children's internalizing problems: Do self-control and parent-teacher meetings matter?

- **Fragile Families:** to adjust standard errors and correctly estimate variance, **jackknife replication** variance estimation was used.

- **Examples of deliverables:**

- Brief reports:
 - <http://www.fragilefamilies.princeton.edu/briefs2.asp>
 - Invited talk (available upon request)

- **The Friendship Project:** All analyses were conducted using structural equation modeling (SEM) with **maximum likelihood estimation** with robust chi-squares and standard errors (MLR). This approach has been shown to adequately address non-normality and unmodeled heterogeneity (Muthén & Asparouhov, 2002; Yuan & Bentler, 2000).

- ML was used instead of restricted maximum likelihood (REML), which is the default in PROC MIXED in SAS given that the models in the present study included fixed effects. REML was not appropriate as it only allows for comparison of nested models that differ in their random effects (Snijders & Bosker, 1999).

- **Examples of deliverables:**

- Wang, J. M., Rubin, K. H., Laursen, B., Booth-LaForce, C., & Rose-Krasnor, L. (2013). Preference-for-solitude and adjustment difficulties in early and late adolescence. *Journal of Clinical Child and Adolescent Psychology*. 42(6), 834-842. doi:10.1080/15374416.2013.
 - Wang, J. M. (2014). Preference-for-solitude, friendship support, and internalizing difficulties during early adolescence in the USA and China. University of Maryland: Unpublished dissertation.

- **BIAS REDUCTION TECHNIQUES:**

- **The Friendship Project:** I examined the normal distribution and discovered that scores on the predictor were skewed. I applied **natural log transformation** to normalize the distribution. I then

conducted separate analyses with the untransformed and transformed data to examine whether there were significant differences in results.

- **Examples of deliverables:**
 - Wang, J., McDonald, K. L., Rubin, K. H., & Laursen, B. (2012). Peer rejection as a social antecedent to rejection sensitivity in youth: The role of relational valuation. *Personality and Individual Differences*, 53(7), 939-942.
- **The Chinese Youth Study:** On average, 0 to 8% of data were missing. Missing data was considered with maximum likelihood estimation with robust chi-squares and standard errors (MLR). In MLR, estimates of standard errors and test statistics are corrected for the degree of data non-normality. MLR provides data-model fit indices and test statistics adjusting for non-normality via the Satorra-Bentler (S-B) correction. Using MLR with S-B test statistics have been shown to adequately address non-normality and unmodeled heterogeneity (Muthén & Asparouhov, 2002; Yuan & Bentler, 2000).
 - **Examples of deliverables:**
 - Wang, J. M. (2014). Preference-for-solitude, friendship support, and internalizing difficulties during early adolescence in the USA and China. University of Maryland: Unpublished dissertation.
- **Early Childhood Longitudinal Study (ECLS):** We examined the normality distributions of measures and discovered that some of the variables were skewed. We applied a **reciprocal transformation** in order to transform its distribution to approximate normality.
 - We also created school-level aggregates for each of the child-level variables by calculating the mean of each variable by school. These aggregates were included in HLM (hierarchical linear modeling) analyses depending on the centering decisions for each child-level variable (more details are available upon request).
 - **Examples of deliverables:**
 - Wang, J. M. & Wimsatt, M. (under review). Young children's internalizing problems: Do self-control and parent-teacher meetings matter?

OTHER NOTES:

- Proficiency in multiple enterprise software (e.g., SPSS, SAS, Mplus, HLM, LISREL) and open-source software (e.g., Python, R, Gephi)
- Experience with scripting and coding (Python)
- References available upon request