5-1-2016

The Inter-rater Reliability of the Psychopathy Checklist-Revised in Practical Field Settings

Yuko Matsushima
Southern Illinois University Carbondale, yuko.mtsm@gmail.com

THE INTER-RATER RELIABILITY OF THE PSYCHOPATHY CHECKLIST-REVISED

IN PRACTICAL FIELD SETTINGS

by

Yuko Matsushima

B.A., International Christian University, 2005
M.A., Senshu University, 2007

AThesis
Submitted in Partial Fulfillment of the Requirements for the
Master of Arts Degree

Department of Criminology and Criminal Justice
in the Graduate School
Southern Illinois University Carbondale
May 2016

THESIS APPROVAL

THE INTER-RATER RELIABILITY OF THE PSYCHOPATHY CHECKLIST-REVISED IN PRACTICAL FIELD SETTINGS

By

Yuko Matsushima

A Thesis Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Art

in the field of Criminology and Criminal Justice

Approved by:

Dr. Daryl G. Kroner, Chair

Dr. George W. Burruss

Dr. Jeffrey Nowacki

Graduate School
Southern Illinois University Carbondale
December, 7, 2015

AN ABSTRACT OF THE THESIS OF

YUKO MATSUSHIMA, for the Masters of Arts degree in Criminology and Criminal Justice, presented on December, 7 (Date of Defense), 2015, at Southern Illinois University Carbondale

TITLE: The Inter-Rater Reliability of the Psychopathy Checklist-Revised in Practical Field Settings

MAJOR PROFESSIR: Dr. Daryl G. Kroner

This paper examined the inter-rater reliability of psychological assessments in practical field with 42 inmates' PCL-R scores. As results, this study showed similar ICC and SEM values to those from PCL-manual. Concerning PCL-R structure, factor 2 showed higher ICC value than factor 1, and facet 4 showed higher ICC value than facet 1, 2, or 3. Especially, facet 2 showed low ICC value. Those are consistent with previous studies. However, ICC yielded by factor 2 only and both factor 1 and 2 showed similar ICC values. Considering theoretical and clinical aspects, it was recommendable to use PCL-R total score as risk assessment, though interpreting facet 2 requires cautions. Concerning to rater's characteristics, the most influential factor to keep the PCL-R reliability was conducting it on regular basis, rather than licensed status. It was difficult to examine whether or not singed-off contribute to maintain sufficient reliability due to small sample size. In regression model, all rater related variables were not significantly correlated to PCL-R score change between two assessment occasions. PCL-R scores at Time 1 was moderately and negatively correlated to PCL-R score change. This indicated natural regression toward the mean. It is desirable to conduct additional study after obtaining more sample and rater related information, such as clinical experience. Additionally, it requires a consideration to apply findings in this study to female psychopathic subjects. As a policy implication, it is recommendable for personnel division to have psychologists to remain in their psychological work.

ACKNOWLEDGEMENTS

I would like to acknowledge my deepest appreciation to my chair Dr. Daryl G. Kroner, for his dedication to support my project. His comments based on the experienced and insightful knowledge on psychological assessment in forensic area helped me develop a critical discussion into the topic. I also learned a lot throughout his classes for correctional treatment and statistics. I would also like to express my gratitude to the other two committees, Dr. George W. Burruss and Dr. Jefferey Nowacki. Dr. George provided me with critical comments based on his extensive knowledge on criminal justice. Dr. Nowacki provided me with valuables knowledge through his class. Finally, I would like to thank my family for supporting me throughout studying in the U.S. Special thanks to Miki Paskert, Nathan Paskert, and Yurino Kawashima, for supporting my life in Carbondale.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER1

INTRODUCTION

The purpose of this paper is to investigate the inter-rater reliability of the PCL-R in field use. The Hare Psychopathy Checklist-Revised (PCL-R; Hare, 1991, 2003) is a clinical rating scale that is widely used in forensic area (Boccaccini, Turner, & Murrie, 2008; DeMatteo & Edens, 2006; Edens & Petrila, 2006). An increasing number of research has discussed the "field reliability" of the PCL-R recently (Boccaccini, Murrie, Rufino, & Gardner, 2014; Boccaccini, Turner, & Murrie, 2008; Edens, 2006; Levenson, 2004). There would be a distinction between a research study, which has been conducted under ideal conditions, and field use. Filed inter-rater reliability is demonstrated by practitioners who have to perform under limited time and regular work conditions (Wood, Nezworski, & Stejskal, 1996). There are serious questions about the field reliability of the PCL-R (Boccaccini et al., 2008; Murrie et al., 2008). Though the PCL-R itself has been empirically validated, there would be several factors that have affected inter-rater agreement in a field use. The PCL-R has been often used for legal-decision making, such as Sexual Violence Predator trails and parole decisions (Boccaccini, Turner, & Murrie, 2008; DeMatteo & Edens, 2006; Edens & Petrila, 2006). The PCL-R plays an important role as a risk-assessment for legal-decision makings, and sufficient reliability of the PCL-R is important because the result of the assessment affects sentencings. Generally speaking, people simply believe that experiences enhance professional skills. However, some research showed that clinical experiences do not contribute to risk prediction or make it worse (Wlaters, Kroner, DeMatteo, & Locklair, 2014; Elbogen, Mercado, Scalora, and Tomkins, 2002), while training could enhance the accuracy of risk prediction (Walters et al, 2014). If experience does not contribute to the accuracy of assessment, what strategies are helpful to keep assessment reliability? Allard and Faust (2000) recommended a set of behaviors for a scoring a test either double checked or optically scanned and computer scored to prevent error caused by human. This paper focuses on the inter-rater reliability of psychological assessments in filed use. Using the PCL-R scores of inmates, this research will examine the factors affecting on the reliability of psychological assessments.

Before discussing the main study, the reason why I chose this topic for thesis will be covered. The author is a psychologist working in Japanese correctional institutions. In 2006, Japan enacted the Law Concerning Penal Institutions and the Treatment of Sentenced Inmates. For nearly 100 years, Japanese correctional administration had been regulated by the Prison Law enacted in 1908. In this old law, prison work was the center of prison life and nothing was clearly mentioned for rehabilitative treatment. The New Law has clearly stated the enforcement of rehabilitative treatment programs or "guidance for reforms", and those were made mandatory. This is why the new law was big change on Japanese correctional history.

After the New Law was enacted, the Japanese prison system introduced rehabilitation programs. One of the major treatment programs is the sexual offender program. In this program, all sexual offenders are assessed on their recidivism risk and are assigned to an appropriate program based on the Risk-Needs-Responsivity (RNR) principles, which was proposed by Canadian psychologists (Yamamoto, 2012). RNR principles are guidelines for effective correctional treatment (Andrews & Bonta, 2010). Before this program had been introduced, risk level of re-offense had not been considered well. There were not clear criteria how to select program participants. Each instructors were used to choose program participants based on their own judge, requirement from other staff members, or participants' motivations. The sexual offender program informed how important risk assessment is for effective correctional treatment.

In this way, the sexual offender program was the first evidence-based practice in Japanese Correction. Following this program, other rehabilitative programs have been developed recently.

Similarly, Japanese-Juvenile correction also developed and introduced an actuarial risk assessment for juvenile correction at the beginning in 2012. Before that, juvenile delinquents' risk had been assessed based on each practitioner's experience. Thus, risk assessments are becoming more important in Japanese corrections. One main reason for this change is policy evaluation. Recently, it has been required much policy evaluation to indicate the efficiency of government works. Along with this trend, the number of actuarial risk assessments will be increased in Japanese corrections. In actual, adult correction bureau in Japan has been developing actuarial risk assessment for all types of offenders.

The current problem in Japanese work

Though Japan has used some risk assessment tools recently, there has been a concern whether or not the field staff recognize how important risk assessments are to their daily work. For example, the concerned situation can be seen in the previously mentioned sexual offender program. Some of the risk-level criteria are determined by consensual agreement, which occurs at a conference. This evaluation is not based on structured data, and evaluators sometimes use experience-based reasons. For example, some evaluators mention about other raters like this "His scoring is always harsh." Emotional reactions may effect on the risk estimation like this "This case is horrible. This offender must reoffend." Some evaluators will score consistently severe, or with particular offense types might be rated more severe because of stereotypical view. Personal experience-based judge sometimes would not provide good reasoning because personal experiences differ among individuals.

Though Japanese corrections has started to focus on Evidence Based Practice (EBP) (Ministry of Justice Japan, 2012: Yamamoto & Matsushima, 2010a: Yamamoto & Matsushima, 2010b: Yuma, Kanazawa, Inotsume, & Matsushima, 2014), many workers have not considered the importance of risk assessments in earnest. If an assessment system does not work properly, classification of program participants would be incorrect, and each participant could not receive the best treatment for them. If an assessment for each case is not conducted appropriately, the result of program evaluation by using those assessment result would get a wrong conclusion. The wider meaning of program indicates the sequential of assessment, treatment, and evaluation of both process and outcome (Rossi, Lipsey, & Freeman, 2003). Generally, people tend to focus on only treatment and the outcome. However, for achieving a treatment goal, the process, which means how a treatment is delivered, should be attended to, and treatment participants should be assigned to a proper treatment by assessment with sufficient reliability. Program evaluations contribute to improve programs and better practice. Based on RNR principles, program participants should be assigned to appropriate programs that match their risk level. It can be said that assessment is the first and essential step of effective treatment programs. The reliability of assessment is essential for successful treatment. This paper will discuss the reliability of psychological assessment and how this can contribute to evidence-based practice.

CHAPTER 2

PSYCHOMETRICS AND RELIABILITY

1. Risk assessment tools

Clinical assessment is a way of evaluating clients' physical, behavioral, and psychological conditions to provide treatment plans (Kroner, Mills, Gray, & Talbert, 2011). There are many forms of clinical assessment: interviews, self-reported questionnaires, neurological and biological tests, and so on. In correctional settings, clinical assessment provides useful information about offenders' clinical problems such as mental health, suicide risk, and violence risk to correctional staff (Kroner, Mills, Gray, & Talbert, 2011). An example of clinical assessment is the Personality Assessment Inventory (PAI). In contrast to clinical assessments, examples of risk assessment instruments are the Level of Service Inventory-Revised (LSI-R), Historical Clinical Risk Management-20 (HCR-20), and STATIC-99. While clinical assessment provides information for treatment plans, risk assessment is useful for estimation of re-offending.

In 1990, Andrews, Bonta and Hoge proposed three general principles of classification for effective correctional treatment: the Risk, Need, and Responsivity (RNR) principles (Andrews, Bonta, & Hoge, 1990; Andrews & Bonta, 2006). The principles has been widely recognized in correctional assessment. The first risk principle is important in two aspects. First, criminal behaviors can be predicted. Second, the first risk principle proposed the matching levels of treatment services to the risk level of the offenders. This matching provides the bridge between assessment and effective treatment. To reduce recidivism, higher risk offender should be assigned in more intensive treatment. As resources for treatment are limited, proper classification is beneficial for appropriate distribution of the resources. The criminogenic need principle refers dynamic and changeable treatment targets that are related to recidivism factors. For example, history of antisocial behavior is helpful for prediction, but that is not changeable because it is a past history. Antisocial cognitions, which include attitudes, values, beliefs, rationalizations, and a personal identity favorable to crime are changeable through treatment. Those can be treatment targets. If the purpose of treatment is reduction of recidivism, then the treatment target should be criminogenic need factors. The general responsivity principle covers the style and mode of delivering treatment programs. The ability and learning style of offenders should be considered to let their program be effective (Andrews & Bonta, 2006). Generally, offenders show lower IQ than average in the society, and many of them have also discrepancy factors relating IQ. Some are not good at absorbing information through reading or listening. Some are not good at abstract reasoning. In this way, all three principles are necessary for effective program delivering. Among three principles, assessment of recidivism risk is essential for the first risk principle. Reliability of assessment, which is the topic of this thesis, is key for the application of the first principle.

The evolution of assessment for correctional treatment could be summarized as follows (Andrews & Bonta, 2006). In the first generation, each professional judges offender's risk in an unstructured manner based on their own professional experience. These judgement are not empirically validated. The second generation tests, including STATIC-99, are empirically validated. However, these are limited to provide useful information for treatment because the second generation tests consists of mostly static factors such as criminal and vocational history and these could not be treatment targets. Almost all second risk assessment tools are not based on theoretical background. They do not provide criminogenic relative information. The third generation tools incorporate theoretically-based criminogenic needs. The LSI-R asks both risk and need factors. However, empirical research and practice in the real world are different. For practitioners, the useful information is how to manage their cases. Criminogenic needs provides information about what should be targeted instead of how to do so. The fourth generation tools focus on the linkage between assessment and case management. To discuss reliable and practical assessment, a solely statistical approach is not sufficient as shown in the steps from the second generation.

Psychological construction and theoretical consideration are required for clinical utility. Describing the detail later, the PCL-R was constructed based on theoretical explanations for psychopathy (Hare, 2003), and much research has reported predictive validity of the PCL-R for recidivism (Hanson and Morton-Bourgon, 2005: Haws, Boccaccini, Murrie, 2013), though the PCL-R was not designed to be an actuarial risk assessment.

2. Classical test theory - errors in assessment

To discuss reliability of psychological assessment, this study will use classical test theory. Classical test theory assumes that each individual has a true score, however, there are always errors in measurement due to multiple reasons when conducting psychometric tests (Kline, 2000). Classical test theory is concerned with the random errors. Classical test theory is a basic theory of test construction for educational and psychological tests. Psychometric tests result in only observed scores, and observed scores are the sum of true scores and some error (random or systematic errors).

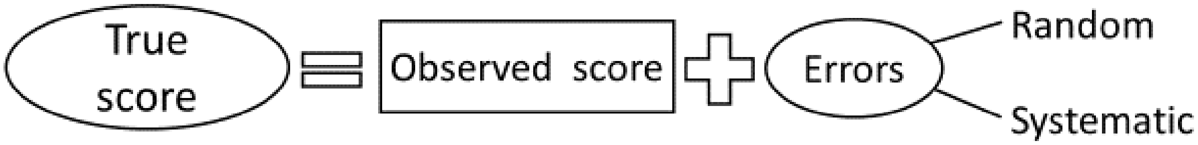

Figure 1
Basic concept of classical test theory

To have a valid true score, it is necessary to decrease the proportion of error variance. Random errors in measurement are caused many reasons, such as moods of examinees, poorly constructed test items which may lead inadequate understanding of the test, insufficient or inappropriate test direction, etc. In addition, errors can be caused by raters. Raters is the main concern of this study.

3. Reliability issues in rated tools

Rating may vary depending on raters' characteristics like personality traits, interview style, training, experience, their background or individual value (Boccaccini, Murrie, Rufino, & Gardner, 2014). When conducting assessments, the acquired scores contain the possible range of errors, or errors of measurement. There is always a certain degree of error and noise because of misunderstanding the true intent of the items, poor administration of test, the changing moods of clients, and so on (Gary, 2009). However, if there is quite large measurement error, test conductors could not discriminate between true scores and errors. Well-developed tests are more likely to have less measurement errors (systematic error), or error fluctuation (random error). That is, it is important to minimize systematic error to keep reliability of assessment.

There are two factors related to the degree of errors in a psychological test (Gary, 2009). The first one comes from natural traits in human performance. The second is the nature of psychology. Psychological tests need to be imprecise to assess "soft" objects. Psychological latent factors such as emotion or mood usually show more variation than concrete concepts like human weight and heights. Because the human mind is not observable, it must be assessed through various latent aspects. For instance, questions measuring anger might ask "are you easily irritated?" or "How often do you get angry?" Anger may be assessed through a variety of similar survey items.

There are four methods to confirm the reliability of the test: internal consistency[1], test-

[1] Internal consistency represents the homogeneity of the test item. This shows the degree to which each item is consistent with others. Those items measure the same constructs. The common way to evaluate internal consistency is either the split-half reliability or Cronbach's alpha. In split-half reliability approach, a test is split in half, the halves are then evaluated through correlation. Similar items should result in a high correlation coefficient. Cronbach's alpha is a measure of the average correlations among all related items.

retest reliability[2], interrater reliability, and parallel forms reliability[3] (Myers & Winters, 2002: Gary, 2009)[4]. Interrater reliability shows the agreement, or concordance between multiple informants. This reliability is mostly discussed for the clinical assessments requiring an interview. Among four types of reliability, inter-rater reliability is a practical issue, while the other reliability are mainly discussed through test construction processes.

Well-developed tests are both reliable and valid (Bachman & Schutt, 2014, pp.90-91). Reliability is "a measure of reliable when it yields consistent scores or observations of a given phenomenon on different occasions (p.88)." Reliability is a prerequisite of the tests, however it is not a sufficient condition for validity. Validity is required for "a test truly measures what it claims (Gravetter & Wallnau, 2014, p.459). Well-developed tests accurately measure the theoretical concepts. In a reliable but invalidated test, respondents answer to the items consistently, however the answers are consistently misleading. If a risk assessment does not have sufficient reliability and validity, the decision based on the risk assessment would be wrong. Legal decision has a large impact on both offenders and victims, and other related people. It is crucial to maintain sufficient reliability of risk assessment to make a right decision.

The main focus of this paper will be interrater reliability. As stated above, most of the risk assessment tools in correctional setting have been validated. However, it is individual raters that use those assessment tools. If individual raters use the tools inaccurately, the empirical validation of the tools will be compromised.

4. PCL-R

[2] Test-retest reliability discusses the stability, whether a test is stable over time. This reliability is especially important for repeated measures, such as in case of treatment progress assessment. A correlation of two administrations is one of the ways to examine stability.

[3] Parallel-forms reliability is concordance between similar forms of tests. This is useful and practical to developing different versions of the same test.

This study will use inmates' Psychopathy Checklist-Revised scores (PCL-R; Hare, 2003) to examine reliability of psychological assessment. The PCL-R is the most common psychopathy scale for measuring antisocial personality (Appendix A). The PCL-R is also well used in forensic areas for risk-assessment to predict recidivism of violent and sexual offenders. For example, it is used in the Sexual Violence Predator (SVP) trials in some states (Edens & Petrila, 2006). The assessment result have a large impact on the trials.

Hare's group recommended to get training to conduct the PCL-R, though the training is not mandatory. They hold a course for practitioners because conducting the PCL-R requires to know the PCL-R well. The test requires 90 to 120 minutes for interviewing and 60 minutes for collateral reviewing administration time.

PCL-R has 20 items, and those are rated by both interviewing subjects and reviewing their file information. It is possible to score the PCL-R through only file reviewing, however interviewing is desirable for more accurate assessment. Each item is three-item Likert scale rated from 0 to 2 (0 = does not apply, 1 = somewhat applies, 2 = definitely applies). The total score ranges from 0 to 40. The PCL-R shows two factors, each with two facets. Factor1 consists of both facet 1 and facet 2: Facet 1 is interpersonal, facet 2 is affective. Factor 2 consists of both facet 3 and facet 4: Facet 3 is lifestyle, facet 4 antisocial. There are the other two items which are not included in any facets. Those are promiscuous sexual behavior and many short-term marital relationships.

PCL-R was used the total of 10,896 offenders and forensic psychiatric patients to calculate percentiles and T-scores. Those sample were from male offenders, male forensic psychiatric patients, and female offenders. Details of samples and the descriptive statistics can be seen in the Hare PCL-R manual (Hare, 2003).

5. Reliability issues with the PCL-R

There are some negative aspects about the PCL-R, though this is widely used in forensic area. First, the definitions of psychopathy is arguable (Cooke, Michie, & Hart, 2006). The concept of psychopathy originally comes from Cleckley's (1941) classic work on "psychopathy", which originally described persons with antisocial personality disorder, and the conceptualization of the PCL-R mostly based on Cleckley's Psychopathy (Hare, 1991, p.2). However, some items in the PCL-R ask simply past criminal behaviors (e.i., juvenile delinquency history, revocation of conditional release, and criminal versatility). Psychopathy and serious criminal offending are different concepts, though there do overlap (Van voorhis, & Salisbury, 2013). Some psychopath commit crimes, and some criminals are diagnosed as psychopath. In contrast, some psychopath do not commit crimes, and some criminals are not diagnosed as psychopath. In addition, those historical items are unchangeable, and those would not work well as clinical judgement materials. Those do not inform treatment providers what treatment targets are. As a second problem, though evaluators get a training to conduct the PCL-R, there may be still evaluators' differences (Boccaccini, Murrie, Rufino, & Gardner, 2014; Boccaccini, Turner, & Murrie, 2008; Edens, 2006; Levenson, 2014). Among several psychological tests, the PCL-R requires a higher skill.

CHAPTER 3

WAYS TO MEASURE INTER-RATER RELIABILITY

This paper focus on inter-rater reliability. In this section, ways of measuring inter-rater reliability are discussed statistically. There are various methods to rater agreement (Burry-Stock, Shaw, Laurie, & Chissom. 1996). The easiest and simplest way is joining probability of agreement, but this is the least robust measure. The method can be applied for only nominal data, and does not take into account that agreement may happen solely by chance. It would cause overestimation the level of agreement (Hallgren, 2012). Especially, when the number of categories in ratings is small, the possibility of agreement by chance increases.

1. Kappa

In contrast to joining probability of agreement, Cohen (1960) proposed how to calculate coefficient of agreement for nominal data with considering agreement expected by chance. The method, called “Cohen’s kappa”, is appropriate for calculating two paired rates’ correspondence. The Cohen’s kappa formula is as follows.

$$\kappa = (Po\text{-}Pc) / (1\text{-}Pc)$$

Where, “Po” is the observed agreement, and “Pc” is the expected probability that two rater agreed by chance, which calculating the probabilities of each rater randomly saying each category. As shown in the formula, the probability of agreement by chance is subtracted. Without subtracting the probability of agreement by chance, the probability of agreement would be higher than the actual. As a disadvantage, it is difficult to obtain sufficient κ value when very uncommon or common conditions are assessed (Xu & Lorber, 2014). For uncommon or common phenomenon, the possibility of agreement by chance would be quite large. Based on the Cohen’s kappa, several types of kappa statistics have been proposed. Fleiss (1971) expanded Cohen’s kappa for applying to three or more raters.

2. Interrater Correlation Coefficient (ICC)

For interval and ratio level data, interrater correlation coefficients (ICC) are mostly used to confirm the interrater reliability. ICC can be used to calculate coefficients of agreement by two or over raters. Because there are many forms of ICCs, an appropriate type should be selected for each research purpose (Hallgren, 2012). Though there is not an unified classification of ICCs, Shrout and Fleiss (1979) discussed six forms of ICCs and this is one of the major classifications of ICCs. According them, three dimensions are there to choose the appropriate form of the ICCs; "(a) Is a one-way or two-way analysis of variance (ANOVA) appropriate for the analysis of the reliability study? (b) Are differences between the judges' mean ratings relevant to the reliability of interest? (c) Is the unit of analysis an individual rating or the mean of several ratings? (Shrout and Fleiss, p.420)"

Though there are not unified ways to represent various ICCs, ICC (n, k) would be one of the easiest ways to represent ICCs variations (Shrout, 1979), where n ranges from 1 to 3, and those numbers indicate Case1, Case2, and Case3 of ICCs, and k indicates the number of raters. Case1 is same as One-Way Classification mentioned by Bartko, Case2 indicates Two-way Random Model, and Case3 indicates Two-way Mixed Model (Bartko, 1966). With the number of raters, it is necessary to consider whether single-measures or average-measures (consistency among multiple-raters). Though consistency agreement does not concern that rater A always assigns higher scores than rater B in same manner, this rating difference of the two raters is a serious issue in risk assessment. In sum, there are 3 × 2 patterns of ICCs.

3. Standard Error of Measurement (SEM)

Another way to discuss how ratings vary is standard error of measurement (SEM). SEM is one of the methods to estimate confidence interval of a population's mean (Kline, 2010). Larger SEM values indicate larger variance of scores. Too large variances are undesirable for reliable assessment. As mentioned before, the classical test theory assumes that observed score consist of true score and measurement errors (Kline, 2010). It is impossible to measure true scores directly, and that must be estimated through observed scores. Standard deviation is obtained in one measurement. If a subject was measured a large number of times, a distribution of scores is obtained, and those scores would shape normal distribution. However, in most cases, it is difficult to measure objects many times. SEM is an estimation of SD if measurements are conducted many times and the true score is constant. SEM is related to SD as the definition, and more reliability leads less SEM.

SEM is calculated with the next formula.

$$\text{SEM} = \text{Observed Standard Deviation } \times \sqrt{1-r}$$

*“r” is the test-retest reliability coefficient

For the test-retest reliability, reliability coefficients yielded by ICC will be used in this study. Higher reliable tests show lower SEMs. Theoretically, 68% of cases should be within one SEM unit, 95% of cases should be within two SEMs unit if the variance comes from only random error. This gives an estimate of the amount of error in the test from statistics that are readily available from any test.

In sum, there are many statistical methods to measure reliability: ICC, Kappa, and SEM. Those are related each other because all those measure reliability, but do from different aspects.

CHAPTER 4

LITERATURE REVIEWS

Though the PCL-R has been well used in practice, the theoretical structure of the PCL-R has been still discussed (Bolt, Hare, Vitale, & Newman, 2004; Hare, 2006; Perez, Herrero, Velasco, & Rodrianez-Diza, 2015), and many researchers expressed concerns about the use of the PCL-R in practical fields (Boccaccini, Murrie, Rufino, & Gardner, 2014; Boccaccini, Turner, & Murrie, 2008; Edens, 2006; Levenson, 2014). The PCL-R is a clinical rating scale, and has 2 factors and 4 facets (Hare, 2003). Those factors and facets have showed different features, and there are still continuing discussions for the structure of factors and facets. Regarding to the practical use of the PCL-R, there seem many reasons to question the reliability: practical situations, which are SVP trials and parole decisions, and both rater's and case's characteristics.

Factor and Facet levels of the PCL-R

The PCL-R are structured with two factors and four facets and each factor and facet has a psychologically constructed concepts as shown in the previous chapter and appendix A. To examine the factor and facet level of PCL-R, rater-agreement seemed stronger for Factor 2 (social deviance) than Factor 1 (Interpersonal/Affective) scores, and Facet 4 (Antisocial) score in Factor 2 also showed stronger rater-agreement than the other facets (Edens, Boccaccini, Johnson, & Johnson, 2010; Hare, 2003; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Sturup, Sorman, Fredriksson, Edens, Karlberg, and Kristiansson, 2014). One of the reasons may be because Facet 4 is assessed with criminal record mostly, which is historical and static. For example, Even if anyone counts the number of conviction based on case-file information, it will be the same number. In contrast, evaluation Facet 1 seems to be more affected by each rater's characteristic (Edens et al., 2010: Miller, Kimonis, Otto, Kline, & Waserman, 2012). Moreover, in general, Factor 2 is a stronger predictor of violence and recidivism than Factor 1 (Hawes, Boccaccini, &

Murrie, 2012: Kennealy, Skeem, Walters, & Camp, 2010). Totally, Factor 2, especially Facet 4 seems to have more reliability and to be stronger predictors. Factor 1 would be more influenced by rater's individual differences. It is useful to know in which parts of the PCL-R raters are more likely to assess differently for being cautious in assessing with the PCL-R. Less research conducted in item level analyses because of the difficulty of obtaining item level data. Along with the facet level's discussion, dynamic and changeable items might show more variance among raters.

The PCL-R for SVP trials and preventative detention designation

The PCL-R is frequently used for legal decisions, such as criminal sentencing and parole decision. Especially, risk assessment including the PCL-R plays an important role for sexual violent predator (SVP) case trials in the U.S. (Dematteo et al., 2013: Murrie et al., 2008, 2009: Rufino et al, 2012: Levenson, 2004). It can be said that risk assessment has a great impact on sentencings. However, field reliability of the PCL-R seems poorer than that for research only.

One possible cause of less reliability is the rater's position when he/she assesses cases in trials. Raters seem to have a tendency to assign scores toward the expectations of the party who retained them. In other words, prosecution-appointed raters may score risk significantly higher than defense-appointed raters. This is called partisan allegiance effects (Blais, 2015). Discussing the PCL-R measurements by raters in different position, previously published studies have shown group differences between raters (DeMatteo et al., 2013; Murrie et al., 2009). For example, Dematto et al. (2013) discussed the U.S. SVP cases that prosecution witnesses had reported the average of PCL-R total scores that were on approximately 5 higher than defense witnesses. Rufino, Boccaccini, Hawes, & Murrie (2012) also examined the difference between opposing forensic experts with using Texas Sexually Violent Predator data. In similar to DeMatteo and his colleagues' (2014) findings, the mean score of the PCL-R for the prosecution was higher than for the defense. They compared researchers' scoring with those opposing forensic experts, and they concluded that prosecution side seemed to give higher scores.

In Canada, after violent or sexual offenders are convicted, the prosecutors can request a consideration of preventative detention designation for the offenders. To make the final designation, the legislation requires that at least one risk assessment by an expert to support the judge. This system is similar to SVP trials in the U.S. Blais (2015) examined the degree to which judges rely on expert information in their decision. As a result, judges showed extreme reliance on expert information. Judges and experts did not show statistical difference in ratings of offenders. However, prosecution-retained PCL-R scores were significantly higher than defense-retained PCL-R scores. This was similar to the U.S. SVP trials' situation, though the mean difference in PCL-R score between two opposite sides in the preventative detention designation was smaller than that of the U.S. SVP trials.

Contrasting to those studies in legal decision-makings, Edens, Smith, Cox, DeMatteo, & Sorman (2015) claimed that the reason why many studies of sexual offense cases tended to show lower reliability in the assessment was most of those cases had been assessed in the legal decision processes. Sexual offending cases are more likely to go through adversarial legal proceedings than other type offense cases, and more likely to be assessed with the PCL-R in trials. Because of this, there is more research by using the PCL-R scores of sexual offenders through legal decision processes. Sexual offense cases are more likely to receive partisan allegiance effects. Edens and his colleagues carefully collected the data to avoid the influence of legal decision process, and they showed the result that ICCs for sexual offending cases was higher than those for non-sexual offending cases. This result is controversial compared to the findings in SVP trial cases.

Based on those findings, it could be said that the reason why evaluations of sexual offense cases sometimes show insufficient rater agreement is the situational influence under legal decision process. The data of the current study were collected in correctional facilities and all cases had already sentenced. Trials process would NOT be a reason to detract good rater agreement in this study.

Research VS Applied ratings

In the previous section, SVP trails in the U.S. and preventative detention designation in Canada were discussed. Beyond those specific situations, the reliability seems weaker when applied in the field, compared with research studies (Boccaccini, Murrie, Rufino, & Gardner, 2014). Generally, in the research studies, well-experienced or trained raters conduct the test under ideally controlled environment. In contrast, in the practical fields, there are many raters, some of whom are well-experienced or trained or some are not. Practical fields may have many restrictions to conduct psychometrics test. There would be a time restriction because of heavy caseloads, and raters may have to assess cases even if there were not sufficient case records.

Relating to the applied ratings, there are some discussions that interview do not add the reliability of the PCL-R assessment or may reduce the reliability (McGrath, 2003: Quinsey & Ambtman, 1979: Wong, 1988). Basically, the PCL-R evaluators give a point to each item when they find any information to support the fact. In other word, if information are more available, clients are more likely to receive higher scores. This seems to be one of the reasons that scoring based on both file review and interview is more likely to give higher points than scoring with file review alone. In addition, the reason why interviews may reduce the reliability of psychological assessments is the human involvements (Allard & Faust, 2000: McGrath, 2003: Wong, 1988). For example, Wong (1988) compared the PCL-R scores of obtained by file review alone and both file review and interview. Those scores did not show significant difference. He concluded that rating the PCL-R should be done with only file review if informative files are available. Quinsey & Ambtman (1979) also compared three types of information source rated by either psychiatrists or teachers for a prediction of offender's dangerousness. Scoring with psychiatric assessment showed far less interrater reliabilities than scoring with offense description or life histories among both psychiatrists' and teachers' group. Similar to Wong, they were suspicious the usefulness of psychiatric assessments. In sum, interviewing, which is considered to require a special training and experience, might increase complexities of ratings and reduce reliabilities of assessment. It could be said that a well-determined process for accuracy of scoring is more important than each rater's "professionality".

Rater's characteristics

Raters' issues are essential topics in conducting psychometric tests appropriately. As stated in chapter 2, there are always measurement errors. The reliability of psychometric tests would be damaged because of large measurement errors. Raters would be one of the factors affecting the reliability of assessment. Though situations would be influencing the rating of the PCL-R, raters' individual characteristics also seems to affect the ratings. One of the most important discussions is whether "professionality" enhances the reliability of psychological assessment or not. It would be a difficult to define what professionality is. Generally, people simply believe that training and experience contribute to be a professional. Previous findings relating to this information are as follows.

Concerning the difference of the employee's status, Rocque and Plummer-Beale (2014) conducted the research for the reliability of the LSI-R in the criminal justice practice, and they compared raters from facilities and communities. They assumed that community group would show higher reliability because raters in the community assess the LSI-R on a regular basis. As a result, the ICC (single rater for absolute agreement) value for the facility group was 0.626, and that for the community group was 0.751. Though the result was not statistically significant mainly because of the small sample size, it could be considered the difference of ICC value was quite large.

There is another question whether or not professionality contributes more accurate risk prediction. Though base rate information contributes better clinical prediction, human judges generally focus on each individual case and think the case represents or resembles a particular category, with little consideration to the relative size of that category. It is referred as base rate fallacy or base rate neglect (Pennycook, Trippas, Handley, & Thompson, 2014). It is an important question whether or not trainings and experiences enhance clinical decision's accuracy. It is simply believed that trainings and experiences are essential to reliable clinical decision-making, however this assumption has little empirical support. For example, Walter et al. showed that both experienced and inexperienced judges tended to neglect the base rate in risk prediction (Walters, Kroner, DeMatteo, & Locklair, 2014). Moreover, the experience in field had negative correlation to hit rate of risk prediction. Though the reason why experience and accuracy showed an inverse correlation should be researched more, they discussed training on the use of base rates may be a solution to keep prediction accuracy.

Here is another research discussing the risk assessment and an individual user. Risk assessment tools are empirically well-validated, however, it is evaluators that use the tools, and it affects the tools' usefulness how evaluators perceive the risk factors derived from empirical research. Elbogen, Mercado, Scalora, and Tomkins (2002) examined how clinicians perceive the relevance of research factors in violence risk assessment. Though all factors were recognized as somewhat relevant for violence prediction, most of the clinicians perceived dynamic and behavioral variables were more relevant than research-based factors. Social history variables, early history variables such as early maladjustment and educational history, were perceived less relevant. In their research neither training nor year of clinical experience differentiated perceptions of risk factors. They discussed that the traditional training for clinicians usually does not contain risk assessments, and that it is necessary to instruct clinicians how research factors predict violence. Training is important, and the content should be related to risk assessment for better prediction.

In conclude, the year of experience themselves seems not to relate to prediction's accuracy. On the other hand, a training on risk assessment and the use of base rate for clinical decision could related to the accuracy. To conduct risk assessment requires understanding empirical research findings.

Multiple raters and repeated assessments

As discussed above, each rater has an individual characteristic, and it seems hard to exclude influence of individual characteristics on psychological evaluations. Rating by multiple raters could increase concordance among raters, and multiple rater agreement might be more reliable than single rater's evaluation.

Allard and Faust (2000) recommended a set of behaviors for a scoring a test either double checked or optically scanned and computer scored to prevent error caused by human. They conceptualized this as commitment to accuracy (CTA). In their research, CTA worked even if a test has complicated procedure and structure. Though the tests which they examined were MMPI, Beck Depression Inventory, Spielberger State/Trait Anxiety Inventory, CTA would work for the PCL-R scoring, too. There was a report that the PCL-R reliability assessed by a forensic evaluation team in Sweden, those members were within the same department (Sturup et al., 2014). They evaluated 27 life sentenced prisoners, and they got a higher ICC (A,1) value than those of previous PCL studies assessed by single raters, though the ICC was not so high (.70) for the total score. Rating by a team seems to be effective for a reliable assessment.

In a practical use, the PCL-R are sometimes conducted two or more times for the same cases, and the evaluators are usually different. In those cases, individual differences of raters have much impact on the assessment reliability. Sturup et al. (2014) reported an aggregated mean of difference scores of 4.9 points (SD=5.1) between two times' PCL-R among 27 life sentenced prisoners in Sweden. In their study, 48% of the difference scores were within 2.9, which is SEM value provided by the PCL manual, and 19% of the scores were between 2.9 and 5.8 apart. Those percentages are far less than theoretical estimations (68% of samples should fall within one SEM and 95% should fall within two SEMs, if it is assumed that their study sample has same size of variances as Hare's (2003) data). They suspicious a clinical use of the PCL-R because of the large measurement error. In contrast, some research reported that sufficient interrater reliability between two occasions (Levenson, 2004). As an example, Levenson (2004) compared the two independent evaluators about same SVP cases in Florida, and he reported relatively high ICC of .84. For a practical use of the PCL-R, the tool should show sufficient interrater reliability because many raters sometimes assess same cases independently.

Psychopathy and sexual aggressive behaviors

While rater's characteristics were discussed in a previous sections, what characteristics of cases do spoil assessment reliability? Are there particular types of cases which assessments are difficult and the results vary among raters? Though SVP trials' situations would affect evaluation, there is another question whether psychopathy is related to sexual offenses putting the trial process story aside. First of all, Hare (2003) stated in the manual that the interaction between the PCL-R and sexual deviance supported the usefulness of the PCL-R for sexual offenders' risk. However, there are inconsistency findings for sexual offenders' assessment.

As previous studies, Porter, Brinke and Wilson (2009) researched the relation between the PCL-R scores and offense types in Canadian federal prison. Though child molesters had lower total scores than the other sexual and non-sexual offender groups, the PCL-R total score did not differ significantly among the rest of groups. In short, the PCL-R scores would not related to particular crime types including sexual offenses. However, the PCL-R was reported in meta-analysis studies as one of predictors of sexual recidivism (Hawes, Boccaccini, & Murrie, 2013: Hanson & Morton-Bourgon, 2005), though the predictive power seemed different between research and clinical use, and effects from field studies were moderate (Hawes et al. 2013). Knight and Guay (2006) overviewed the research for psychopathy and sexual coercion and they concluded that there would be a relation. They discussed the relation between psychopathy and sexual coercion in three research areas. Psychopathy seemed to be more likely to be rapists than non-psychopathy. The component of psychopathy, especially the impulsivity and antisocial deviance had predicted sexually coercive behavior in convicted offenders. Similarly, in noncriminal samples, the component of psychopathy predicted rape.

Mostly, psychopathy and sexual aggressive behaviors seem to have relation, and the PCL-R could be said to be a predictor, however the effect was moderate and varied among field studies. And some research showed the PCL-R less reliable, especially in the legal decision-making. Again, this study would provide a new finding for this discussion as one of the findings from correctional facilities.

CHAPTER 5

HYPOTHESES

To address the reliability issues, I propose three hypotheses.

Hypothesis #1

Because it is assumed that field data have larger errors of measurement and show less inter-rater reliability,

1A the SEM of this study will be larger than 2.90, which was found by Hare (2003).

1B ICC for single rater of absolute agreement would be less than the values in the PCL-R manual (Hare, 2003), which is 0.88.

Hypothesis #2

Assessing historical and static information will obtain higher agreement among raters. Facet 4 (Antisocial) can be mostly assessed with case-file information because most items in facet 4 ask past information. Factor 2 consists of both facet 3 and facet 4. Because of this,

2A Regarding the ICC for Factor levels, Factor 2 of the PCL-R will show higher ICC than Factor 1.

2B Similarly, Facet 4 of the PCL-R will show higher ICC than Facet 1, 2, or 3 of the PCL-R.

2C With regard to item level, historical items (juvenile delinquency history, revocation of conditional release, and criminal versatility) will show larger Kappa than the other items.

Hypothesis #3

Assessing the PCL-R regularly will contribute to better reliability. Licensed psychologists will show higher reliability because they pass through requiring for being licensed. Checking scores by multiple-raters would help to maintain the scoring accuracy. Because of this, ICCs for regular assessment conducting, licensed psychologists, and score checking will be as follows.

3A Employees of Correction Service of Canada will have lower ICC values than contract employees because contract employees conduct the PCL-R on regular basis.

3B Licensed Psychologists will show higher ICC values than non-licensed psychologists. (Approximately 80 percent of licensed psychologists have doctoral degree, while similar percentages of non-licensed psychologists have master's degree as their final background. Because the license and academic degree are highly correlated, this study will focus on the license based on the assumption that fulfilling the requirements for licensed more influential factor than academic degrees to be a professional psychologist.)

3C Rating by a pair of non-licensed and licensed psychologists will show higher ICC than rating by a licensed psychologist alone.

CHAPTER 6

METHODOLOGY

Data source

The data came from two federal prisons in Canada. The inmates in this dataset were consecutive admissions to federal custody from June 1995 to August 1996. They had been conducted their first assessment within two months of the inmates' arrival of at an assessment unit. In the dataset of this study, there are 45 inmates assessed two or more times. Although 2 participants had received the fourth time's test, the fourth time's score was not analyzed in this study because of the very small sample size. Three cases were assessed by same raters at time 1 and time 2, and which were excluded from the data analysis after checking the characteristics of those three cases[5]. The final sample was 42 cases, and each case was assessed by different raters. Ten cases were assessed three times, but sample of 10 is not sufficient for analyses. Time 1 and Time 2 scores will be mainly used for analysis.

Twenty-seven participants (64.3%) were incarcerated in Bath Institution, and 15 participants (35.7%) were incarcerated in Pittsburgh Institution. Bath Institution is a medium-security correctional facility, and Pittsburgh Institution is a minimum-security facility. Both are located in Ontario, Canada.

The average time between the first and second conduct was 4 years and 69.9 days (SD=1022.75), and minimum length was 1 year and 31 days and maximum length was 11 years and 357 days. The average period between the second and third conduct was 3 years and 225.6

[5] Two of them were assessed twice, and the rest one was assessed three times. The case A's PCL-R total scores were 10 at time 1, and 12 at time 2. The case B's scores were 26 and 17. The case C's scores were 27, 19, and 28. Case A and B were assessed by CSC employees, and Case C was assessed by contract employees, and all those three raters were licensed psychologists with Ph.D. Final sample was 42 cases at time 1 and time 2, and 10 cases at time 3.

days (SD=893.86), and minimum length was 306 days and maximum length was 9 years and 23 days.

Case Characteristics

Classifying based on the crime motivation, there were 10 sexual offenders (31.3%), and 22 non-sexual offender (68.8%), and 11 cases had missing values. Other case characteristics not relating to this study directly described in Table 1 and footnote[6].

Table 1
Description of Case Characteristics

		n	%
Types of offenses	murder	32	76.2%
	sexual offenses	4	9.5%
	Robbery	3	7.1%
	Assault	1	2.4%
	Arson	1	2.4%
	criminal negligence/major driving	1	2.4%
Life sentenced cases		29	69.0%
DO designation		4	9.5%

Generally, most of offenders are assessed only one time with the PCL-R in Canadian correctional institutions as one of the references for parole decision. The sample in this current study has at least 2 times of the PCL-R scores. There are some possible reasons. The most common reason to be assessed repeatedly would be that they failed to be granted parole decision by parole boards, though it is impossible to know the reason from the current dataset.

There were some missing data, and those were replaced[7]. Prorated total scores (Hare,

[6] Case characteristics: With regard to types of offenses, 32 cases (76.2%) were murder, 4 cases (9.5%) were sexual offenses, and 3 cases (7.1%) were robbery, and there was one case for assault (2.1%), arson (2.1%) and criminal negligence/major driving (2.1%). Life sentenced cases were 29 (87.9%), "DO designation" (judged as dangerous offender through preventative detention designation) cases were 4 (12.1%), and 9 cases were neither life sentenced nor judged as DO designation. Few cases had race/ethnicity.

[7] Missing data: The first PCL-R assessment had nine items missing across the 20 items. The most frequent missing items 11.1 % (n=5) for the item of "Any Short-Term Marital Relationships" and "Juvenile Delinquency". The second PCL-R session had six items missing. The most frequent data missing were 9.1 % (n=4) for the item of "Revocation of Conditional Release". The third PCL-R session had no missing item. The occurrence was not frequent and

2002) were not used for analyses in this current study. Prorated total score is the missing data method for clinical purpose. Instead of prorated total score, the sum score of twenty items that including replaced data as stated earlier was used for data analyses in this study.

Description of Raters

At Time 1, there were 30 raters for 42 cases. Concerning those raters' degree levels, 23 cases were assessed by raters with Ph.D., 2 cases were assessed by raters with Ed.D., 14 cases were assessed by raters with MA, and the raters for the rest of 3 cases were unknown or data missing. 27 cases were assessed by CSC employees, and 15 cases were assessed by contract employees. 27 cases were assessed by licensed psychologists, and 14 cases were assessed by non-licensed psychologists, and 1 cases did not have information. Among those 14 cases assessed by

Table 2
Descriptive Statistics of Rater Groups at each time

	Time 1 (n=42)		Time 2 (n=42)		Time 3 (n=10)	
	n	%	n	%	n	%
CSC employees	27	64.3%	14	33.3%	1	10.0%
contract employees	15	35.7%	28	66.7%	9	90.0%
Ph.D or Ed.D	25	59.5%	27	64.3%	7	70.0%
MA or less	14	33.3%	14	33.3%	3	30.0%
Unknown	3	7.1%	1	2.4%	0	0.0%
licensed	27	64.3%	30	71.4%	7	70.0%
non-licensed	14	33.3%	12	28.6%	3	30.0%
Unknown	1	2.4%	0	0.0%	0	0.0%
signed-off	36	85.7%	42	100.0%	10	100.0%
non-signed off	6	14.3%	0	0.0%	0	0.0%

considered as missing completely at random (MCAR), and the median of nearby points method was used for replacing those missing values. After this procedure, the replaced scores were summed up to calculate total score. Some total score had a dismal, which were rounded up. After doing median of nearby points method, there was still one blank in case No.2 at 17th item at the second conduct. For this missing data, the most frequent value (that was 0) is imputed.

Prorated total scores (Hare, 2002) were not used for analyses in this current study. Prorated total score is the missing data method for clinical purpose. Instead of prorated total score, the sum score of twenty items that including replaced data as stated earlier was used for data analyses in this study.

non-licensed psychologists, 4 cases were not assessed without licensed psychologists' checking. Among 27 cases assessed by CSC employees, 18 cases were assessed by licensed psychologists, and 6 cases were assessed by non-licensed psychologist, and 3 cases did not have relative information.

For Time 2, there were 26 raters. For this assessment, 27 cases were assessed by raters with Ph.D., 12 cases were assessed by raters with MA, and 2 cases were assessed by raters with BA or less, and the rater's degree for one case was unknown. Fourteen cases were assessed by CSC employees, and 28 cases were assessed by contract employees, which was a big different from Time 1. Thirty cases were assessed by licensed psychologists, and 12 cases were assessed by non-licensed psychologists.

For Time 3, there were 6 raters. Concerning those rater's degree levels, 7 cases were assessed by raters with Ph.D., 1 cases were assessed by raters with MA, and 2 cases were assessed by raters with BA or less. 1 cases were assessed by CSC employees, and 9 cases were assessed by contract employees. 7 cases were assessed by licensed psychologists, and 3 cases were assessed by non-licensed psychologists.

There were ten raters who participated in both Time 1 and Time 2. Among those raters, there were four raters who participated in all conducts (Time 1, 2, and 3).

Rational for hypothesis 1

This study use data from a field practice, and it is assumed that there is a distinction between a research study and a field use of assessment. Field inter-rater reliability (Wood, Nezworski, & Stejskal, 1996) is demonstrated by practitioners who have to perform under limited time schedule and their regular work conditions, while experimental research is usually conducted under ideal condition by well-trained or experienced raters. In the practical fields, some of raters are well experienced or trained or some are not. Practical fields often have many restrictions to conduct psychometrics test. There would be a time restriction because of heavy caseloads, and raters sometimes have to assess cases even if there were not sufficient case records.

Because of this, it is hypothesized that inter-rater reliability of this study will lower than that of PCL-R manual. Similarly, SEM of this study will larger than that of PCL-R manual.

Rational for hypothesis 2

As discussed in literature review section, it was reported that Factor 2 showed higher reliability than Factor1, and Facet 4 showed higher reliability than Facet 1, 2, and 3 (Edens, Boccaccini, Johnson, & Johnson, 2010; Hare, 2003; Miller, Kimonis, Otto, Kline, & Wasserman, 2012; Sturup et al., 2014). Along with this, this study expects to obtain same result.

Rational for hypothesis 3

Concerning a professional psychologist, academic background and psychologist's license are important factors to think about their professionality (DeMatteo, Marczyk, Krauss, & Burl, 2009; Bedi, Klubben, & Barker, 2012). As discussed in literature review session, being licensed usually requires doctoral level background or equivalent experience.

The current study has four types of information about raters; employment status (full time employment in Correction Service of Canada or contract employment), academic degree, licensed or non-licensed psychologist, signed off or not (Signed off means that a licensed psychologist checks a PCL-R report conducted by another psychologist).

The current study has two raters' groups, which are Correction Service of Canada (CSC) employees and contract employees. In CSC, the contract employees have been employed for assessment work, and they assess cases as part of their regular work. If conducting assessment as regular work contributed more reliable assessment, the contract employee's group will show higher reliability than CSC staff. On the other hands, assessing serious offenders may require special experiences in correctional fields (Van voorhis & Salisbury, 2013), because psychopathic offenders may attempt to charm, deceive people, and manipulate them. Discussing psychopathic manipulation, Seto and Barbaree (1999) also reported that sexual offenders who participated treatment programs and got the most positive evaluations by instructors also showed the highest PCL-R scores and the highest recidivism rates. This shows the difficulty of evaluating psychopathic inmates. At this point, CSC employees would have more advantageous experiences to treat troublesome offenders without involving with their psychopathic behaviors. It is hypothesized that contract employees will show higher inter-rater reliability than CSC employees do.

With regard to academic degree and psychologist registration, approximately 60 % of raters in this study had a doctoral degree, and the rests have a master's degree or less. Similarly, approximately 60 % of raters at Time 1 and approximately 70% of raters at Time 2 were licensed psychologists, and the rests were non-licensed psychologists. The requirements to be licensed varies among jurisdictions in Canada, and each province and territory has specific licensing requirements for working as psychological practitioners in a given province or territory (Canadian Psychological Association, 2015). Typically, doctoral level is desirable for the license and the master's level psychologists required additional supervised experience before becoming licensed.

Though the psychologist in criminal justice field is a large group in Canadian Psychological Association (CPA), it is reasonable to assume their background education is clinical or counseling psychology through their doctoral course (Bedi, Klubben, & Barker, 2012). CPA accredits preferable training programs to be a professional psychologist, though the license itself is approved by each provincial or territorial bodies (CPA, 2015). The number of law-psychology and forensic psychology has increased within this 20 years, however it is desirable to get a general and foundational level of competence, and specialization in a particular area of forensic area would be postdoctoral level (DeMatteo, Marczyk, Krauss, & Burl, 2009). The doctoral program of clinical and counseling psychology vary among universities (Bedi, Klubben, & Barker, 2012), and to be a registered psychologist usually requires additional supervised experience and examination after getting academic degrees. In this way, it could be assumed that license is more influential than an academic degree to be a professional psychologist. Because having a doctoral degree is related to being licensed, this study will assume that fulfilling the requirements for licensed will capture having academic degrees.

With regard to signed-off, it is reasonable that the signed-off system will contribute to assessment accuracy. Allard and Faust (2000) recommended a set of behaviors for scoring a test of either double checked or optically scanned and computer scored to prevent human error. Psychologists using a signed-off system can check basic mistakes or discuss case assessment.

Analyses Strategy

For hypothesis #1, SEM will be calculated for each time's assessment. SEM will be calculated with the next formula.

$$\text{SEM} = \text{SD} \times \sqrt{1 - r} \qquad {}^{*}r \text{ is the test-retest reliability coefficient.}$$

Larger SEM indicates less reliability. Along with the literature review, it is expected that the SEM values in this study will be larger than SEM value provided by PCL-R manual (Hare, 2003).

The time length between Time 1 and Time 2, and between Time 2 and Time 3 varied. The relationship with the length of period between each assessment will be examined because the time length are varied among cases. If difference scores became larger when the time length between two assessments was longer, there would be an increased correlation, and this would indicate that time length between two assessments might be related to score changes.

ICC(A,1) and Cronbach's alpha will be calculated for the total, Factor, and Facet scores. ICC(A,1), which is a two-way random effect model for a single rater with absolute agreement, is appropriate to examine rater agreement between individual raters (McGraw & Wong, 1996). "A" means absolute agreement and "1" means a single rater. ICC(A,1) assumes that there is a systematic factor of variance associated with both raters and subjects. True score consists of observed score and measurement error. Two way random model with inmates and raters as random factors provides variance component estimates for inmates, raters, and the interaction of inmates and raters.

$x_{ij} = \mu + T_i + J_j + I_{ij} + E_{ij}$

Where, (x: observed score, μ: true score, Ti: the effect of inmate i, Jj: the effect of rater j, Iij: residual effect of the inmate i and rater j, Eij: measurement error)

$$\text{ICC(A,1)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_J^2 + \sigma_I^2 + \sigma_E^2}$$

Based on the hypothesis #1, ICC(A,1) for the total score between two occasions will be lower than those of Hare's (2003). With regard to hypothesis #2, Factor 2 and Facet 4 will show higher ICC(A,1) and coefficient α. For item level, Kappa will be calculated.

In regard to hypothesis #3, there are some available information about raters' characteristics (CSC employee or not, licensed psychologist or not, and sign off or not). ICC will be conducted for each variable and rater group using Time 1 scores.

Next, regression (Ordinal Least Squares) will be conducted using rater related variables as independent variables (CSC employee or not, licensed psychologist or not, signed off or not) and the difference scores between Time 1 and Time 2 as a dependent variable. Because not all raters assessed PCL-R both Time 1 and Time 2, raters' information at Time 1 and Time 2 were different. Each regression analysis for Time 1 and Time 2 will be conducted using each rater related data. If the two regression show similar model, it can be interpreted the model predict score change regardless of raters shifting. PCL-R total score at Time 1, whether sexual offenders or not, and institutional information will be entered as control variables in the regression (Table 3).

Mathematically, how large a score differs from the average have an influence on the degree of change because of regression toward to mean. This is a reason that Time 1 scores will be considered as control variable. There is considerable discussion how psychopathy relates to sexual offenses (Knight and Guay, 2006; Hare, 2003; Hawes, Boccaccini, & Murrie, 2013; Morton-Bourgon, 2005; Porter, Brinke, & Wilson, 2009), and it is unreasonable to ignore sexual offense variable. Sexual offenders were coded as 0, non-sexual offenders were coded as 1. Concerning institutional information, "local groups" might affect an accuracy of scoring (McGrath, 2003). Local groups might have local rules which deviate from test manuals, or all the group member might misunderstand how to use and interpret a test through a study group within the limited group member. It sometimes happens that misunderstandings are spread among members through a small study group. Especially, correctional facilities are much closed environment because of the nature of the correction. It would be reasonable to assume that "local" variances are easy to occur in correctional facilities. Inmates in Bath prison were coded as 0, and Inmates in Pittsburgh were coded as 1.

Table 3

Variables in Regression (Ordinal Least Squares)

	Time 1	Time 2
Independent Variables	n	n
Employment Status		
0 CSC employees	27	14
1 contract employees	15	28
Licensed		
0 licensed psychologists	27	30
1 non-licensed psychologists	14	12
unknown	1	
signed off for non-licensed psychologists		
0 signed-off	6	0
1 non-signed-off	36	42
Dependent Variables		
the difference scores of the PCL-R total scores between Time1 and Time2	42	42
ranges from -40 to 40		
Control Variables		
PCL-R total scores at Time1	42	42
ranges from 0 to 40		
Sexual motivation for offenses		
0 sexual motivation	10	10
1 non-sexual motivation	22	22
missing value	10	10
Institutions		
0 Bath	27	27
1 Pittsburgh	15	15

CHAPTER 7

RESULTS

Correlation between the time lengths between assessments and difference scores

Descriptive statistics of PCL-R scores was shown in Table 4. The time lengths between Time 1 and Time 2 assessments and the degrees of difference score was not significantly correlated (*r* = -.038, *n.s.*). Considering the purpose of this analysis, difference scores were converted to absolute value. The correlation coefficient between the time lengths and the absolute value of difference scores was not also significant (*r* = .074, *n.s.*).

Table4
PCL-R Scores at Time 1, 2, and 3

PCL-R	Time 1 (n=42)		Time 2 (n=42)		Time 3 (n=10)	
	M	SD	M	SD	M	SD
Total	19.8	7.81	20.0	7.16	21.1	6.26
Factor 1	7.1	4.16	7.7	4.06	6.8	2.94
Facet 1	3.2	2.32	3.5	2.23	2.6	1.84
Facet 2	3.9	2.35	4.2	2.33	4.2	1.55
Factor 2	9.1	3.92	9.4	4.04	11.1	3.25
Facet 3	4.7	2.63	5.1	2.43	6.1	2.38
Facet 4	5.5	2.52	5.4	2.86	6.2	2.15

Factor1 includes item 1, 2, 4, 5, 6, 7, 8 and 16. Facet1 includes item 1, 2, 4 and 5. Facet2 includes item 6, 7, 8, and 16. Factor2 includes item 3, 9, 10, 12, 13, 14, 15, 18, and 19. Facet3 includes item 3, 9, 13, 14, and 15. Facet 4 includes item 10, 12, 8, 19, and 20.

Hypothesis 1A to 2B

Concerning hypothesis 1B, ICC(A,1) of the PCL-R total scores for Time 1 and Time 2 was 0.85 (95% Confidence interval [.74-92]). The ICC value was almost same as those from PCL-R manual (0.86 for male offenders, and 0.88 for male forensic patients). The hypothesis 1B was not supported.

Concerning hypothesis 2A, ICC (A,1) of the PCL-R Factor 1 for Time 1 and Time 2 was 0.61 [0.38-0.77]. ICC (A,1) of the Factor 2 for Time 1 and Time 2 was 0.82 [0.69-0.90] .For facet level, which is hypothesis 2B, ICC (A,1) of the Facet 1 for Time 1 and Time 2 was 0.70 [0.51-0.83]. ICC (A,1) of the Facet 2 for Time 1 and Time 2 was 0.50 [0.24-0.70]. ICC (A,1) of the Facet 3 for Time 1 and Time 2 was 0.73 [0.55-0.84]. ICC (A,1) of the Facet 4 for Time 1 and Time 2 was 0.81 [0.68-0.89]. In similar to ICCs for Time 1 and Time 2, ICCs among three occasions were calculated as shown in Table 5. Though the ICC values among three occasions were a little bit lower than those for Time 1 and Time 2, similar results were obtained. That is to say, hypotheses 2A and 2B were supported.

Table 5
ICCs for PC-R Total, Factors and Facets scores

PCL-R	Time 1-2 (n=42)		Time 1,2, & 3 (n=10)	
	ICC	95% C.I.	ICC	95% C.I.
Total	0.85	[.74-.92]	0.87	[.68-.96]
Factor 1	0.61	[.38-.77]	0.57	[.18-.85]
Facet 1	0.70	[.51-.83]	0.64	[.30-.88]
Facet 2	0.50	[.24-.70]	0.46	[.05-.80]
Factor 2	0.82	[.69-.90]	0.78	[.52-.93]
Facet 3	0.73	[.55-.84]	0.61	[.25-.87]
Facet 4	0.81	[.68-.89]	0.83	[.61-.95]

Concerning Hypothesis 1A, for yielding the SEM value, pooled standard deviations of Time 1 and Time 2 was used. Formula of pooled SD was shown below.

$$\text{SD}_{\text{pooled}} = \sqrt{\frac{\{(n1-1)S1^2+(n2-1)S2^2\}}{n1+n2-2}}$$

Where, S1=7.81, S2=7.15, n1=n2=42, the $\text{SD}_{\text{pooled}}$ was,

$$= \sqrt{\frac{(42-1)\times 7.81^2+(42-1)\times 7.15^2}{42+42-2}}$$

$$= 7.491$$

Formula of SEM was as follows.

$$\text{SEM} = \text{SD} \times \sqrt{1-r} \qquad *r \text{ is the test-retest reliability coefficient.}$$

ICC of the PCL-R total score between Time 1 and Time 2 was 0.85. The SEM of this study is,

$$\text{SEM} = 7.49 \times \sqrt{1-0.85}$$

$$=2.90$$

This was same as the SEM value of 2.90 which provided by the PCL-R manual (Hare, 2003). Hypothesis 1A, which expected the SEM values of this study is larger than that of PCL-R manual, was not supported.

Hypothesis 2C

With regard to items level between Time 1 and Time 2, Kappa coefficients were calculated as shown in Table 6. Because there were only 10 cases at Time 3, Kappa coefficients between Time 2 and Time 3 were not calculated. As shown in Table 4, the size of Kappa coefficients varied among items from 0.18 to 0.69. Item 9 "Parasitic lifestyle" showed the highest value (0.69), and item 18 "Juvenile Delinquency" showed the second highest value (0.61). In contrast, item 15 "Irresponsibility" and item 16 " Failure to accept responsibility for own actions" showed the lowest values (0.18). Historical items seem to be more likely to obtain higher ICC values, while items related to affective aspects seem to be more likely to obtain less ICC values. However, the items which showed the third highest ICC value (0.53) are the next three: “1. Glibness/Superficial Charm”, “2. Grandiose sense of self-worth”, and “19. Revocation of Conditional Release”. Hypothesis 2C expected that items to be scored mainly by official record showed higher ICC values. However, item 20 “Criminal Versatility” did not obtain high ICC value (0.42), and some items which are from interpersonal aspects (item1 and item2) showed high ICC values.

Table 6

PCL-R Item Scores and Kappa between Time 1 and Time 2

	first conduct		second conduct		third conduct		Kappa between
Item	Mean	SD	Mean	SD	Mean	SD	T1 and T2
1. Glibness/Superficial Charm	0.8	0.82	0.8	0.76	0.7	0.67	0.53
2. Grandiose Sense of Self Worth	0.6	0.62	0.7	0.73	0.6	0.52	0.53
3. Need for Stimulation/Proneness to Boredom	1.1	0.76	1.1	0.75	1.5	0.53	0.41
4. Pathological Lying	0.8	0.79	0.7	0.71	0.2	0.42	0.32
5. Conning/Manipulative	1.0	0.73	1.2	0.66	1.1	0.99	0.28
6. Lack of Remorse or Guilt	1.1	0.72	1.1	0.78	1.0	0.47	0.34
7. Shallow Affect	0.8	0.79	0.9	0.76	0.9	0.32	0.31
8. Callous/Lack of Empathy	0.9	0.75	1.2	0.74	1.2	0.79	0.49
9. Parasitic Lifestyle	0.7	0.73	0.7	0.70	0.9	0.74	0.69
10. Poor Behavioral Controls	1.2	0.82	1.1	0.79	1.6	0.52	0.32
11. Promiscuous Sexual Behavior	1.1	0.82	1.2	0.84	1.4	0.84	0.38
12. Early Behavioral Problems	1.0	0.84	1.0	0.83	1.3	0.82	0.43
13. Lack of Realistic, Long-Term Goals	0.8	0.69	0.8	0.79	0.9	0.88	0.44
14. Impulsivity	1.1	0.71	1.3	0.65	1.5	0.71	0.42
15. Irresponsibility	1.0	0.76	1.1	0.71	1.3	0.82	0.18
16. Failure to Accept Responsibility for Own Actic	1.0	0.75	1.0	0.73	1.1	0.74	0.18
17. Many Short-Term Marital Relationships	0.5	0.62	0.5	0.71	0.6	0.70	0.36
18. Juvenile Delinquency	1.0	0.82	0.9	0.91	0.9	0.88	0.61
19. Revocation of Conditional Release	1.2	0.91	1.3	0.91	1.2	0.92	0.53
20. Criminal Versatility	1.1	0.81	1.1	0.83	1.2	0.92	0.42

Hypothesis 3A

There were 11 cases assessed by CSC employees at both Time 1 and Time 2. There were 12 cases assessed by contract employees at both Time 1 and Time 2. The other 19 cases were assessed by a CSC employee and a contract employee. ICC values obtained from those three groups was shown in Table 7. Totally, contract employees showed higher ICCs, though ranges of confidential interval were too large to discuss statistical significant differences as shown in the table because of the small sample size. Contract employees showed higher ICCs on total, both factor 1 and 2, facet 1, and facet 2. Assessment by two CSC employees and assessment by a CSC employee and a contract employee showed similar ICC values. It may be because scoring by CSC employees have larger variance. Relating hypotheses 2A and 2B, factor 2 and facet 4 showed high ICC values.

Table 7
ICCs based on employment status at Time 1 and Time 2

PCL-R	CSC employees (n=11)		Contract employees (n=12)		CSC & Contract employees (n=19)	
	ICC	95% C.I.	ICC	95% C.I.	ICC	95% C.I.
Total	0.86	[.55 - .96]	0.91	[.71 - .97]	0.82	[.59 - .92]
Factor 1	0.55	[-.07 - .86]	0.78	[.41 - .93]	0.52	[.10 - .78]
Facet 1	0.66	[.12 - .90]	0.79	[.45 - .94]	0.62	[.25 - .83]
Facet 2	0.52	[-.12 - .85]	0.74	[.34 - .92]	0.36	[-.11 - .70]
Factor 2	0.74	[.28 - .92]	0.88	[.64 - .96]	0.81	[.58 - .92]
Facet 3	0.69	[.21 - .90]	0.69	[.22 - .90]	0.78	[.49 - .91]
Facet 4	0.78	[.36 - .94]	0.86	[.60 - .96]	0.79	[.53 - .92]

Hypothesis 3B

There were 22 cases assessed by licensed psychologists at both Time 1 and Time 2. There were 6 cases assessed by non-licensed psychologists at both Time 1 and Time 2. Most of the cases are signed off for the scoring. It was assumed that signed-off was beneficial for maintain the reliability. However, because of the sample size, both non-signed off cases and signed-off cases were included together to analyze licensed status. The number of non-signed off cases was 6 at Time1, and 4 cases of them were assessed by non-licensed psychologists, 1 case was assessed by a licensed-psychologist, and 1 case did not have an information about the rater's license. All cases were signed-off at Time2 and Time3.

As shown in Table 8, non-licensed psychologists showed very high ICC values. However, because the sample size was too small, it would be considered as reference values.

In contrast, there were relatively more of licensed psychologists. It would not be reasonable to compare licensed psychologists group and non-licensed psychologists group directly due to unbalanced sample size. ICC values calculated with 42 raters' data (Table 5) was used for comparison. As a result, all ICC values yielded with licensed psychologist group were lower than those with 42 raters. Though it is difficult to discuss statistical difference, hypothesis 3B seems not to be supported. If anything, non-licensed psychologists seems to show higher ICC value than licensed psychologists.

Table 8

ICCs Based on Licensed Situation at Time 1 and Time 2 (and ICCs Yielded with Total Data)

PCL-R	non-licensed (n=6)		licensed (n=22)		T1-T2 (n=42)	
	ICC	95% C.I.	ICC	95% C.I.	ICC	95% C.I.
Total	0.94	[.61 - .99]	0.69	[.39 - .86]	0.85	[.74-.92]
Factor 1	0.92	[.54 - .99]	0.37	[-.05 - .67]	0.61	[.38-.77]
Facet 1	0.96	[.76 - .99]	0.60	[.25 - .81]	0.70	[.51-.83]
Facet 2	0.62	[-.39 - .94]	0.25	[-.18 - .60]	0.50	[.24-.70]
Factor 2	0.90	[.46 - .99]	0.75	[.49 - .89]	0.82	[.69-.90]
Facet 3	0.86	[.28 - .98]	0.59	[.23 - .81]	0.73	[.55-.84]
Facet 4	0.80	[.10 - .97]	0.76	[.51 - .90]	0.81	[.68-.89]

Hypothesis 3-C

All 42 cases were assessed with signed off at Time2. Six cases were assessed without signed-off at Time1. For examining the hypothesis about signed-off, the conditions at Time1 were focused. There were 22 cases assessed by both licensed-psychologists at Time 1 and Time 2, and only one case was assessed by licensed-psychologists without signed-off at Time 1. There were only six cases assessed by both non-licensed psychologists at Time 1 and Time 2, and the sample size was too small to conduct ICC. Because of this, 14 cases which were assessed by non-licensed psychologist at Time 1 were used for this hypothesis, though 8 cases among those were assessed by licensed psychologists at time2 (Table 9).

Based on the hypothesis, it was expected that scoring by a non-licensed psychologist without signed-off shows lower reliability. However, scoring without signed-off showed higher ICC value on Facet 2, while scoring with signed-off showed higher ICC values on Facet 1 and Factor 2. The inconsistency result may be because of small sample size. It would be difficult to discuss the hypothesis about signed off with this sample size.

Table 9
ICCs Based on Signed-off among Non-licensed Psychologist

PCL-R	non-signed-off (n=4)		signed-off (n=10)	
	ICC	95% C.I.	ICC	95% C.I.
Total	0.91	[.20 - .99]	0.92	[.70 - .98]
Factor 1	0.80	[-.02 - .99]	0.81	[.43 - .95]
Facet 1	0.47	[.58 - .95]	0.83	[.49 - .96]
Facet 2	0.89	[-.23 - .99]	0.74	[.24 - .93]
Factor 2	0.88	[.08 - .99]	0.95	[.82 - .99]
Facet 3	0.92	[.35 - .99]	0.91	[.68 - .98]
Facet 4	0.88	[.06 - .99]	0.86	[.54 - .96]

Regression

Before conducting ordinal least squares regression, Pearson's correlations of related variables were examined. Each Time 1 and Time 2 data were analyzed separately. As expected, PCL-R total scores and difference scores from Time 2 and Time 1 were significantly correlated at both Time 1 and Time 2 as shown in Table 10 and Table 11 (Time 1, *r* = -.42, *p*<.01; Time 2, *r* =

-.42, *p*<.01). Licensed situation and employment status were correlated at Time 1 (*r* = -.31, *p*<.05). Licensed status of raters and inmates' institution were correlated at Time 2 (*r* = .42, *p*<.01). Because all cases were signed off at Time 2, this variable was not included in correlation and ordinal least squares using Time 2 data.

Table 10
Correlations of Variables in Regression at Time 1

	n	1	2	3	4	5	6
1 Difference scores from Time 2 to Time 1	42	-					
2 Employment status	42	-.042	-				
3 Licensed status	41	-.013	.307*	-			
4 Signed Off	42	.083	-.304	.284	-		
5 PCL-R total at time1	42	-.417**	-.106	-.265	-.053	-	
6 Sexual offender	32	-.170	-.104	.065	-.217	.125	-
7 Current Institution	42	-.115	.170	-.120	-.020	-.274	-.018

**p*<.05. ** *p*<.01.

Table 11
Correlations of Variables in Regression at Time 2

	n	1	2	3	4	5
1 Difference scores from Time 2 to Time 1	42	-				
2 Employment status	42	-.008	-			
3 Licensed status	42	-.064	.224	-		
4 PCL-R total at time1	42	-.417**	-.120	.259	-	
5 Sexual offender	32	-.170	-.035	.080	.125	-
6 Current Institution	42	-.115	.422**	-.251	-.274	-.018

** *p*<.01.

Table 12

Ordinary Least Squares Using Time 1 Data to Predict Score Change between Time 1 and Time 2

Independent Variables	Unstandardized Coefficient (beta)	Standardized Coefficient (b)	
Employment status	.469	.055	
Licensed situation	-2.075	-.239	
Signed off	1.402	.119	
PCL-R total score at time1	-.289	-.543	**
Sexual offender	-.540	-.061	
Current institution	-2.574	-.300	
(constant)	7.536		
Adjusted R^2	10.20%		

*** *p*<.001

Table 13

Ordinary Least Squares Using Time 2 Data to Predict Score Change between Time 1 and Time 2

Independent Variables	Unstandarized Coefficient (beta)	Standarized Coeffcient (b)	
Employment status	.411	.047	
Licensed situation	-.104	-.011	
PCL-R total score at time1	-.249	-.467	**
Sexual offender	-1.012	-.115	
Current institution	-2.296	-.268	
(constant)	5.69		
Adjusted R^2	10.00%		

*** *p*<.001

The ordinary least squares regression model using Time 1 data was not statistically significant as shown Table 12 (F(6, 24) = 1.571, *n.s.*). 10.2% of variation was explained in the dependent variables (see Table 12). The following regression equation was obtained: Y=5.69 + (0.411) (Employment Status) + (-0.104) (Licensed situation) + (-0.249) (PCL-R total score at Time 1) + (-1.012) (Sexual offender) + (-2.296) (Current institution). PCL-R scores ranged from 0 to 40, while the other independent variables were dichotomous. Among independent variables, PCL-R total score at Time 1 showed significantly moderate and negative correlation (*r* = -0.54).

For every increase PCL-R score at Time 1, the degree of difference score from Time 2 and Time 1 decrease by 0.289. After coefficients standardized, PCL-R total score at Time 1 had the most effect on score change between Time 1 and Time 2. Because PCL-R total scores at Time 1 and difference scores from Time 2 to Time 1 were moderately and negatively correlated in Pearson's correlation, the regression results is within expectation.

As shown in Table 13, The ordinary least squares regression model using Time 2 data was not statistically significant, either (F(5,26) =1.693, *n.s.*). 10.00% of variation was explained in the dependent variable (see Table 13). The model was similar to the other model which using Time 1 data. As same as the other model, PCL-R total score at Time 1 showed moderate and negative correlation (*r* = -0.47). For every increase PCL-R score at Time 1, the degree of difference score from Time 2 and Time 1 decrease by 0.249. In sum, commonly in each ordinary least squares, only PCL-R total score at Time 1 significantly affected on the degree of score change, though the regression model was not significant. Considering PCL-R total score at Time 1 was negatively correlated to difference scores from Time 2 to Time 1 in Pearson's correlation, there would be natural regression toward the mean.

CHAPTER 8

DISCUSSION

The PCL-R is widely used in forensic practice in much of the U.S. Researchers have been concerned about the PCL-R use in the practical field (Boccaccini, Murrie, Rufino, & Gardner, 2014; Boccaccini, Turner, & Murrie, 2008; Edens, 2006; Levenson, 2014). However, this study showed the PCL-R reliability in correctional facilities was similar to PCL-R manual (Hare, 2003). Measurement error in this study was not larger than in the PCL-R manual. Factor 2, especially Facet 4 in Factor2, showed larger ICC values than Factor 1 and Facets 1, 2, or 3. Regarding rater's characteristics, contract employee, who conducted assessment on regular basis, showed larger inter-rater reliability than CSC employee, who had variety of work including administration. This study did not find a variable impacting the score change between two successive assessments. Of note, the first conduct PCL-R scores were negatively correlated to the score change, possibly due to natural regression toward the mean.

Hypothesis 1, which hypothesized that the use of psychological assessment in practice would have less reliability, was not supported. The results indicated that the PCL-R use in correctional facilities had similar reliability as the PCL-R manual. Majority of the PCL-R research was conducted with SVP trials data, having partisan allegiance effects (Blais, 2015). Partisan allegiance effects reflect raters having a tendency to assign scores toward the expectations of the party who retained them. In contrast, raters in correctional facilities do not need to consider expectations of their party because inmates have already convicted and sentenced. That is to say, there would be not partisan allegiance effects in correctional facilities. There is minimal conflict of interest regarding PCL-R scores in correction. This is one possible reason why this study showed sufficient reliability of the PCL-R assessment. If so, use of the PCL-R in correctional settings would not have reliability problems.

With regard to rater characteristics, the most influential factor to predict PCL-R reliability was conducting assessment on regular basis. Hypothesis 3A, which expected that employees of Correction Service of Canada will have lower ICC values than contract employees because contract employees conduct the PCL-R on regular basis, was supported. This study examined signed-off process for maintaining assessment reliability (hypothesis 3C) because singed-off would reduce human error (Allard Faust, 2000). However these findings were unclear in this study due to small sample size.

The time length between Time 1 and Time 2 assessment and the difference score was not significantly correlated. The sentencing period would not be expected to effect on the PCL-R score's change, based on the assumption that change has linearly increased. Based on the psychopathy definition, psychopathic personality traits do not change for a few years because the nature of personality is quite stable (American Psychiatric Association, 2013). This result is reasonable considering to the nature of personality trait. It can be assumed that the psychopathic trait itself should not change, and the PCL-R score change would result from other factors.

With regard to hypothesis 1B, SEM value of this study was similar as the SEM value provided by the PCL-R manual, which was 2.9 (Hare, 2003). Previous research by Sturup and his colleagues using field data showed less than 68% of their sample had PCL-R total scores at Time 2 that fall within ±2.90 points of the PCL-R total scores at Time 2, though they did not calculated SEM value itself (Sturup et al., 2014). If Sturup and his colleagues calculated the SEM values, those would be larger than 2.90 because it can be assumed that the variances were larger than that of PCL-R manual provided. The sample of Sturup's study was life-sentenced prisoners in Sweden. Contrary to the current study finding, Sturup and his colleagues showed concern of the PCL-R use in correctional settings.

Regarding factor and facet levels of discussion of the PCL-R, hypotheses 2A and 2B were supported. That is to say, Factor 2 showed higher ICC value than Factor1, and Facet 4 showed higher ICC value than Facet 1, 2, or 3. Especially, Facet 2 showed lower ICC value than other facets. Some researchers claimed only Factor 2 should be used for risk prediction because of insufficient reliability of Factor 1 (Hawes et al., 2013). They stated that Factor 1, which contains emotional characteristics items, is more likely to be impacted by an individual rater’s values. However, the PCL-R total score yielded similar ICC values for both Factor 2 and the PCL-R total score. Taking into account the clinical and theoretical consideration, it would be the best to use the PCL-R total scores for risk assessment rather than Factor 2 only. Regarding facet level, Facet 4, which can be assessed mostly by file information, showed higher reliability. In contrast, Facet 2, which assesses affective aspect, seems to have less reliability. In addition, Kappa values varied among PCL-R items. Some historical items showed higher Kappa coefficients than other items, though some items in interpersonal aspects showed relatively high Kappa coefficients. In sum, it is recommendable to use PCL-R total score as risk assessment, and interpreting Facet 2 would require some caution.

Regarding raters' characteristics, contract employees showed larger ICC than CSC employees, and hypothesis 3A was supported. As discussed in the literature review, regular assessment would contribute to maintain a reliability of the assessment (Rocque and Plummer-Beale, 2014). Contrary to hypothesis 3B, licensed psychologists seemed to show lower reliability. With regard to signed-off, inconsistent results were obtained depending on each Factor and Facet. The sample for this analysis was 4 and 10, which is quite small. It was difficult to discuss effectiveness of signed-off with this sample size. In sum, regular assessment would be more important for maintain the reliability than a rater having a license.

Through regression analysis, PCL-R total score at Time1 affected on the score changes to the next assessment. Subject who obtained higher PCL-R scores at Time 1 were more likely to get lower scores at Time 2 comparing to their first scores. It was already examined that time length did not effect on PCL-R score changes, and this would represent stability of personality trait. Because of this, the score reduction from Time 1 to Time 2 would be reflected of natural regression toward the mean. Beside regression toward the mean, other variables did not show statistical significance. The regression model was not significant, and explained approximately 10% of variances. There may be other variables which may explain assessment reliability.

Limitations

The limitations of this study are as follows. First, the sample of this study was small. Statistical power is related to sample size. Statistical power is the probability of observing significant result if a true difference exists (Lachin, 1981). Small sample size is related to low statistical power. There is a clear method of power analysis for t-test, but not for rater's agreement statistics (Ip, Wasserman, & Barkin, 2012). It is difficult to determine how many subjects were enough for the current study. However, it is obvious that confidence interval of each ICC value was quite large, and it was almost impossible to discuss statistically significant differences. Ideally, it is desirable to have a larger sample. However, most of the research using ICC statistics have small sample because ICC are commonly used for pilot study of clinical trials or clinical judge (Ip et al., 2012). Actually, most of the previous research about field reliability of PCL-R had similar sample size to the current study (DeMatteo, et al., 2013: Edens, Boccacini, & Johnson, 2010; Levenson, 2014: Ruffino, Boccaccini, Hawes, & Murrie, 2012: Sturup et al., 2013). It might be unavoidable to discuss ICC without statistical significance.

Second, this study did not have female data. Female psychopathy seems to have different features comparing to male psychopathy (Hicks, Vaidyanathan, & Patrick, 2010; Krammer, Krueger, & Hicks, 2008). Female psychopathy may be different in two ways: mean-level differences and differences in structure. First, female psychopathy shows lower PCL-R scores than males (Krammer et al., 2008). Second, female psychopathy are more likely to show psychopathological maladjustment (Hicks et al., 2010), though the psychopathic structures between men and women are mostly similar (Kennealy, Hicks, & Patrick, 2007). Because there may be gender differences, and it requires a consideration to apply findings in this study to female subjects.

Third, it is desirable to discuss how rater's length of experience in practical field relates to the accuracy of PCL-R scoring. For example, Elbogen et al. (2002) and Walters et al. (2014) discussed the relation between clinical experience and accuracy of assessment. The current study did not have information regarding the length of work. Clinical experience information would deepen the discussion of how clinical experience relates to assessment reliability.

Implication for future studies

As discussed in previous section, this study had some limitations. It is desirable to gather further information. As noted above, length of clinical experience would be beneficial to examine. Another variable is work tasks in the fields. In Japan, government employees have to do many types of work even though some are employed as specialists including psychologists. A few psychologists in Japanese correction spend a longer time for administrative task than with psychological work. In addition, there are various types of work among psychological work. These include assessment, individual counseling, and group treatment. Even the start point is same when being employed as psychologists, the later experiences of those psychologists vary a lot. To discuss how clinical experience impacts on reliability of psychological assessment, it would be very important to know the variety of work experience for each rater.

Relating to what types of work, the frequency of conducing assessment would also vary among raters. This current study indicated that regular basis assessment contributed to keep assessment reliability. As a practical question, how much frequency can be said as a regular conduct? For instance, it would be helpful to know how many times a rater conducts assessment in a month for examining what “regular” indicates.

Training is important to ensure reliability of psychological assessment (Walters et al, 2014). Some psychologists attend workshops voluntary for psychological work. Training information would make it clear how trainings effect on raters to maintain reliability of assessment.

As a policy implication, it is recommendable for personnel division to have psychologists to remain in their psychological work. Even if a personnel division employs licensed psychologists, they may diminish their professional knowledge and skills if they have not used them for a long time. Though it is expected that trainings would support to keep the knowledge and skills, the current study did not examined this topic. It requires further research to discuss whether training helps to keep reliability of psychological assessments or not, and if so, what kinds of trainings would work.

Conclusion

This study examined inter-rater reliability of the PCL-R in correctional settings. Though previous studies have been concerned the PCL-R use in practice, this study demonstrated similar reliability as stated the PCL-R manual. Conducting PCL-R on regular basis seemed to contribute to keep the inter-rater reliability. This study did not find a major factor impacting on score change between two successive assessments except natural regression toward the mean. This study mainly suffered from small sample size. It is desirable to have a larger sample and conduct further research with other variables relating to rater’s characteristics, such as clinical experience.

REFERENCES

Allard, G., & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment*, *7*(2), 119–129.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, (5th ed)*. Arlington, VA: American Psychiatric Publishing.

Andrews, D. A. & Bonta, J. (2010). *The psychology of criminal conduct (5th ed)*. Albany, NY: Lexis Nexis Anderson publisher.

Andrews. D.A., Bonta, J., & Hoge, R. D. (1990). Classification for effective rehabilitation: rediscovering psychology. *Criminal Justice and Behavior*, 17. 19-52.

Andrews, D.A. & Bonta, J (1995). *The Level of Service Inventory- Revised*. Tronto: MultiHealth Systems.

Bachman, R. & Schutt, R. K. (2014). *The Practice of Research in Criminology and Criminal Justice.* Thousand Oaks, CA: Sage Publication.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*. 19, 3-11.

Bedi, R. P., Klubben, L. M., & Barker, G. T. (2012). Counselling vs. clinical: A comparison of psychology doctoral programs in Canada. *Canadian Psychology/Psychologie Canadienne*, *53*(3), 238–253.

Blais, J. (2015). Preventative detention decisions: Reliance on expert assessments and evidence of partisan allegiance within the Canadian context. *Behavioral Science & the Law, 33*(2), 74-91.

Boccaccini, M. T., Murrie, D. C., Rufino, K. A, & Gardner, B. O. (2014). Evaluator differences in Psychopathy Checklist-Revised factor and facet scores. *Law and Human Behavior*, *38*(4), 337–345.

Boccaccini, M. T., Turner, D. B., & Murrie, D. C. (2008). Do some evaluators report consistently higher or lower PCL-R scores than others? Finding from a statewide sample of sexually violence predator evaluations. *Psychology, Public Policy, and Law,* 14, 262-283.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the psychopathy checklist--revised. *Psychological Assessment*, *16*(2), 155–168.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater agreement indexes for performance assessment. *Educational and Psychological Measurement, 56*(2), 251-262.

Canadian Psychologist Association. (2015). What is accreditation? Retrieved from http://www.cpa.ca/accreditation/whatis/

Canadian Psychological Association. (2015). Provincial and Territorial Licensing Requirements. Retrieved from http://www.cpa.ca/accreditation/PTlicensingrequirements/

Cleckley, H. (1941). *The mask of Sanity*. Saint Louis: C.V.Mosby.

Cohen, J. (1960). A coefficient of agreement of nominal scales. *Educational Psychological Measurement,* 20, 37-46.

Cooke, D. J., Michie, C.& Hart, S. D. (2006). Facets of Clinical Psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp.512-532). New York, NY: Guilford Press.

Cronbach, L. J. (1951). "Coefficient alpha and the internal structure of test". *Psychometrika,* 16 (3): 297–334.

DeMatteo, D., & Edens, J. F. (2006). The role and relevance of the Psychopathy Checklist-Revised in court: A case law survey of U.S. courts (1991-2004). *Psychology, Public Policy, and Law*, 12, 214-241.

DeMatteo, D., Edens, J. F., Galloway. M., Cox. J., & Smith. S. T. (2014). The role and reliability of the Psychopathy Checklist- Revised in U.S. Sexually violent predator evaluation: A case law survey. *Law and Human Behavior,* 38(3), 248-255.

DeMatteo, D., Marczyk, G., Krauss, D. a., & Burl, J. (2009). Educational and training models in forensic psychology. *Training and Education in Professional Psychology*, *3*(3), 184–191.

Douglas, K. S., Hart, S. D., Webster, C. D., & Belfrage, H. (2013). *HCR-20V3: Assessing risk of violence – User guide*. Burnaby, Canada: Mental Health, Law, and Policy Institute, Simon Fraser University.

Edens, J. F., & Boccacini, M. T., & Johnson, D. W. (2010). Inter-rater reliability of the PCL-R total and factor scores among psychopathic sex offenders: Are personality features more prone to disagreement than behavioral features? *Behavioral Sciences and the Law,* 28, 106-119.

Edens, J. F., & Petrila, J. (2006). Legal and ethical issues in the assessment and treatment of psychopathy. In C. Patrick (Ed.), *Handbook of psychopathy* (pp.573-588). New York, NY: Guilford Press.

Elbogen, E. B., Mercado, C. C., Scalora, M. J., & Tomkins, A. J. (2002). Perceived Relevance of Factors for Violence Risk Assessment: A Survey of Clinicians. *International Journal of Forensic Mental Health*, *1*(1), 37–47. doi:10.1080/14999013.2002.10471159

Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters" in *Psychological Bulletin*. 76(5), 378–382.

Gary, G. M. (2009). *Handbook of Psychological Assessment (5th ed)*.Hoboken: Wiley.

Gravetter, F. J., & Wallnau, L. B. (2014). *Essentials of Statistics for the Behavioral Sciences (8th ed.).* Belmont, CA: Wadsworth.

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Hanson, R. K., & Morton-Bourgon, K. E. (2005). The characteristics of persistent sexual offenders: a meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73*(6), 1154–1163.

Hanson, R. K., & Thornton, D. (2000). Improving risk assessments for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24*(1), 119–136.

Hare, R. D. (1991). *Manual for the Hare Psychopathy Checklist-Revised*. Tronto, ON, Canada: Multi-Health Systems.

Hare, R**.** D. (2003). *Manual for the Hare Psychopathy Checklist-Revised (2nd ed.).*Toronto, ON, Canada: Multi-Health Systems.

Hawes, S. W., Boccaccini, M. T., & Murrie, D. C. (2013). Psychopathy and the combination of psychopathy and sexual deviance as predictors of sexual recidivism: Meta-analytic findings using the Psychopathy Checklist-Revised. *Psychological Assessment*, 25(1), 233-243.

Hicks, B. M., Vaidyanathan, U., & Patrick, C. J. (2010). Validating female psychopathy subtypes: Differences in personality, antisocial and violent behavior, substance abuse, trauma, and mental health. *Personality Disorder: Theory, Research, and Treatment*, 1(1), 38-57.

Ip, E. H., Wasserman, R., & Barkin, S. (2013). Comparison of intraclass correlation coefficient estimates and standard errors between using cross-sectional and repeated measurement data: The Safety Check Cluster Randomized Trial. *Contemp Clin Trials,*,32(2), 225-232.

Kennealy, P. J., Skeem, J. L., Walters,G. D., & Camp, J. (2010). Do core interpersonal and affective traits of PCL-R psychopathy interact with antisocial behavior and disinhibition to predict violence? *Psychological Assessment*, 22(3), 569-580.

Knight, R. A., & Cuay, J. (2006). The role of psychopathy in sexual coercion against women. In C. Patrick (Ed.), *Handbook of psychopathy* (pp.512-532). New York, NY: Guilford Press.

Kline, P. (2000). *The Handbook of Psychological Testing (2nd ed.).* New York, NY: Routledge.

Kramer, M. D., Krueger, R. F., & Hicks, B. M. (2008). The role of internalizing and externalizing liability factors in accounting for gender differences in the prevalence of common psychopathological syndromes. *Psychological Medicine*, 38, 51-61.

Kroner, D. G., Mills, J. F., Gray, A., & Talbert, K. O. N. (2011). Clinical assessment in correctional settings. In T. Fagan J. & R. Ax K. (Eds.), *Correctional mental health: From theory to best practice* (pp. 79–102). Thousand Oaks: Sage.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled clinical trials, 2,* 93-113.

Levenson, J. S. (2004). Reliability of sexually violent predator civil commitment criteria in Florida. *Law and Human Behavior*, *28*(4), 357–368.

McGrath, R. E. (2003).Enhancing accuracy in observational test scoring: the comprehensive system as a case example. *Journal of Personality Assessment, 81*(1), 104-110.

Miller, C. S., Kimonis, E. R., Otto, R. K., Kline, S. M., & Waserman, A. L. (2012). Reliability of risk assessment measures used in sexually violent predator proceedings. *Psychological Assessment*, 24(4), 944-953.

Ministry of Justice Japan. (2012). The report of the evaluation of sexual offenders program in Japanese prisons with recidivism data. Retrieved from http://www.moj.go.jp/content/000105287.pdf

Myers, K., & Winters, N. C. (2002). Ten-years review of rating scales. Ⅰ: Overview of scale functioning, psychometric properties, and selection. *J Am Acad Child Adolescent Psychiatry*, *41*, 2.

Quinsey, V. L., & Ambtman, R. (1979). Variables affecting psychiatrists' and teachers' assessments of the dangerousness of mentally ill offenders. *Journal of Consulting and Clinical Psychology*, *47*(2), 353-362.

Rossi, P.H., Lipsey, M.W., & Freeman, H.E. (2003). *Evaluation-Systematic Approach (7th ed.).* CA: Sage Publication.

Rufino, K.,Boccaccini, M.,Hawes, S., and Murrie, D. (2012) When experts disagreed, who was correct? A comparison of PCL-R scores from independent raters and opposing forensic experts. *Law and Human Behavior,* 36(6), 527-537.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.

Seto, M. C., & Barbaree, H. E. (1999). Psychopathy, treatment behavior and sex offender recidivism. Journal of Interpersonal Violence, 14, 1235-1248.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428.

Van voorhis, P. & Salisbury, E. J. (2013). *Correctional Counseling and Rehabilitation (8th ed)*, Routledge.

Walters, G., Kroner, D., Dematteo, D., & Locklair, B. (2014). The impact of base rate utilization and clinical experience on the accuracy of judgements made with the HCR-20. *Journal of Forensic Psychology Practice*, 14, 288-301.

Wong, S. (1988). Is Hare's psychopathy checklist reliable without the interview? *Psychological Reports*, 62. 931-934.

Wood, J., Nezworski, M., & Stejskal, M. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science*, 7, 3-10.

Yamamoto, M. (2012). The presence and future of sexual offenders program (1). *Keisei*, 122(9), 56-64.

Yamamoto, M., and Matsushima, Y. (2012). The presence and future of sexual offenders program (2). *Keisei*, 123(10), 86-95.

Yamamoto, M., and Matsushima, Y. (2012). The presence and future of sexual offenders program (3). *Keisei*, 124(11), 70-79.

Yuma,Y.,Kanazawa,Y.,Inotsume,Y.& Matsushima,Y. (2014). Evaluating the effects of the rehabilitative program for sex offenders in Japanese prisons. *Japanese Criminal Psychology*, 52, 6-7.

APPENDICES

Appendix A

Factors and Facets Structure of the PCL-R

Table 1

Factors and Facet Structure of the PCL-R

Factor 1	**Factor 2**
Facet 1: Interpersonal 1. Glibness/Superficial Charm 2. Grandiose Sense of Self Worth 4. Pathological Lying 5. Conning/Manipulative	Facet 3: Lifestyle 3. Need for Stimulation/Proneness to Boredom 9. Parasitic Lifestyle 13. Lack of Realistic, Long-Term Goals 14.Implusivity 15.Irresponsibility
Facet 2: Affective 6. Lack of Remorse or Guilt 7. Shallow Affect 8. Callous/Lack of Empathy 16. Failure to Accept Responsibility for Own Actions	Facet 4: Antisocial 10.Poor Behavioral Controls 12. Early Behavioral Problems 18. Juvenile Delinquency 19. Many Short-Term Marital Relationships 20. Criminal Versatility
Other items 11. Promiscuous Sexual Behavior 17. Many Short-Term Marital Relationships	

VITA

Graduate School
Southern Illinois University

Yuko Matsushima

yuko.mtsm@gmail.com

International Christian University
Bachelor of Art, March 2005

Senshu University
Master of Art, March 2007.

Thesis Title:
The Inter-Rater Reliability of the Psychopathy Checklist-Revised in Practical Field Settings

Major Professor: Dr. Daryl G. Kroner