The Canadian Journal of Program Evaluation Vol. 23 No. 1 Pages 87–112
ISSN 0834-1516 Copyright © 2009 Canadian Evaluation Society

RELIABILITY OF THE CANADIAN INCIDENCE STUDY DATA COLLECTION INSTRUMENT

Della Knoke
Independent Consultant
Toronto, Ontario

Nico Trocmé
McGill University
Montreal Québec

Bruce MacLaurin
University of Calgary
Calgary, Alberta

Barbara Fallon
University of Toronto
Toronto, Ontario

Abstract: The Canadian Incidence Study of Reported Child Abuse and Neglect (CIS) provides national estimates of the scope and characteristics of reported child maltreatment. It is the most comprehensive source of information about maltreated children in Canada. The data are used to inform policy and, in secondary analyses, to examine correlates of maltreatment and relationships among case characteristics and service responses. Evaluation of the psychometric properties of the CIS instrument is essential to establish confidence in the data collected and the estimates generated. This study examines the test re-test reliability of the CIS instrument.

Résumé : L'Étude canadienne sur l'incidence (EIC) des signalements de cas de violence et de négligence envers les enfants est la source la plus exhaustive de données nationales sur l'étendue et les caractéristiques des signalements de mauvais traitements. Les données servent à orienter les politiques et constituent une base pour des analyses secondaires de facteurs liés aux mauvais traitements et aux caractéristiques des cas et des services fournis. Une évaluation des qualités psychométriques du questionnaire EIC est essentielle pour établir la fiabilité des données recueillies et des prévisions produites. Cette étude examine la fidélité test-retest du questionnaire EIC.

Corresponding author: Nico Trocmé, Room 300, Wilson Hall, 3506 University Street, Montreal, QC, H3A 2A7; < nico.trocme@mcgill.ca>

INTRODUCTION

The demand to demonstrate the effectiveness of child welfare services is consistently being noted in the professional and research literature (Flynn & Bouchard, 2005; Pecora, 2002; Vandermeulen, Wekerle, & Ylagan, 2005). In Canada, child welfare is a mandatory service that is either directly delivered by government employees or, in some jurisdictions, by independent nonprofit agencies funded by government. Service volumes and the cost of providing child welfare services have increased substantially. For example, in 2005, the province of Ontario estimated that it spends more than $1.1 billion annually on direct child welfare services, more than double the amount spent in the late 1990s (Ontario Ministry of Children and Youth Services, 2005). Similarly, the annual expenses for Alberta Children's Services increased by approximately 80% between 2000/01 and 2006/07 (Alberta Children's Services Annual Report, 2000–2001, 2006–2007). Developing systematic approaches to monitor the changing needs of clients and demands on the child welfare system is critical to evaluate system performance and to meet the increasing demand for accountability in social services (Sieppert, 2005).

There is relatively little information about the outcomes of child welfare intervention in Canada. In a review of the impact of child welfare programs between 1995 and 2005, Flynn and Bouchard (2005) identified only 10 peer-reviewed published studies. Provinces and territories have encountered significant challenges in their attempts to adopt a systematic approach to child welfare outcomes evaluation. These challenges are due, in part, to the needs-driven focus of child welfare practice, the shifting tensions inherent in child welfare legislation and policy, and the definitional and technological challenges of measuring outcomes at provincial and national levels (Trocmé, MacLaurin, & Fallon, 2000). At the 1998 Canadian Roundtable on Child Welfare Outcomes, key stakeholders in policy, practice, and research from across Canada endorsed an incremental outcomes development strategy, including as its first priority examining the utility of systems-based indicators generated from administrative data (Thompson & Fallon, 1999). Harvesting administrative data at a provincial or national level is a critical step in describing the child welfare service trends and the impact of policy. The present study examines the test re-test reliability of a data collection instrument used in a national study to provide a standardized approach to recording child welfare administrative data.

Administrative data may be used to identify service trends and emerging needs to inform program and policy design and adjustment. In many health and social service sectors across Canada, there are no mechanisms to compile and integrate client-level information. Access to client-level data has been an area of particular concern in understanding and evaluating policy and service delivery trends with respect to child welfare services for abused and neglected children. The aggregation of information across jurisdictions is limited by differences in definitions of maltreatment, the structure of administrative systems, and the data elements recorded. In addition, administrative data systems are frequently structured to demonstrate accountability for use of government funds and may not consistently include data elements relevant for other purposes (e.g., English, Marshall, Brummel, & Orme, 1999; Fluke, Yuan, & Edwards, 1999). Data elements pertaining to the clinical characteristics of children and families served may not be included or may be less likely to be complete if these data fields are not mandatory. Concerns about administrative data are not specific to child welfare. For example, a study on access to medical services (e.g., MRI) found that types of information relevant to policy—including patients' presenting symptoms and reason(s) for referral, data elements needed to accurately determine waiting time, and the impact of the intervention upon outcomes—were missing from administrative databases (Iron, Przybysz, & Laupacis, 2003).

The Canadian Incidence Study of Reported Child Abuse and Neglect (CIS) provides a mechanism for compiling and integrating data to respond to a significant information gap in the child welfare sector. The primary objective of the CIS is to provide reliable national estimates of the scope and characteristics of investigated child maltreatment. The CIS data also provide information on the characteristics of families and children served by child welfare systems; service responses to child maltreatment and the data are used to identify disparities among subgroups in service decisions. Social workers directly report case-level information collected during their investigations. The CIS collects the types of information that would be contained in clinical files and administrative databases, but it circumvents the problems noted above by collecting data directly from workers using a standardized tool for recording information about child welfare clients and practice. Access to workers provides a mechanism for collecting information that would otherwise not be available. The utility of these data in monitoring key characteristics and trends in service responses is supported by the completeness of the data, the comprehensive array of characteristics surveyed, and the application of standard definitions and classification criteria. These features have made the CIS the most comprehensive source of information about abused and neglected children and trends in child maltreatment in Canada. Given the value of the CIS in the Canadian child welfare context, ensuring data integrity is critical.

Within the context of program evaluation, the CIS serves some functions of a macro-level needs assessment. Nguyen, Attkinson, and Bottino (1983) suggest that needs assessments provide a “fundamental navigational system for program planning and modification based on continuous assessment of changing community needs” (p. 107). The CIS provides data to profile clients and patterns of service over time. Within each cycle, the CIS provides descriptive information about the incidence of various forms of maltreatment, the characteristics of investigated families and their children, and the service decisions made by workers early in the life of a case, such as referral for counselling services or placing a child in out-of-home care. Across cycles, it identifies changes in service demand and service drivers and in the needs of service recipients. Study findings inform policy development, program planning, and decision-making about how to enhance or modify services to help children who have been maltreated or are at risk of being maltreated.

Two waves of CIS data have been collected from representative samples of agencies from child welfare service areas across Canada, excluding Quebec. In 1998, the first CIS (CIS-98) collected information from a sample of 7,672 reports of suspected child abuse or neglect, selected from a sample of 51 child welfare authorities (Trocmé et al., 2001). The second cycle of data collection (CIS-2003) provided information on a sample of more than 11,000 children from 55 child welfare service areas (Trocmé et al., 2005). In addition to providing national estimates, oversampling was conducted in three jurisdictions (Ontario, Alberta, and the Northwest Territories) and for Aboriginal agencies to provide information about patterns and changes specific to these contexts. A third wave of data collection is underway.

Analyses of national data over the first two cycles documented an 86% increase between 1998 and 2003 in the incidence of child maltreatment investigations and a 125% increase in the rate at which maltreatment was substantiated in Canada (Trocmé et al., 2005). Factors such as expanded definitions of child maltreatment, increased public awareness, increased focus on risk and earlier intervention, and the more systematic investigation of siblings in households under investigation for the maltreatment of one child contributed to these increases. Forms of maltreatment showed differential rates of increase. The incidence of sexual abuse showed no significant change. Increases were evident in all other forms. However, exposure to domestic violence and emotional maltreatment showed the largest increases in the rate of substantiation, with increases of 259% and 276%, respectively. These differential increases indicate changes in the service needs of child welfare clients and in the nature of the demands on the system. The smaller proportions of children who exhibit physical or emotional harm and declines in the proportion of children placed in out-of-home care suggest child welfare services are reaching a broader range of children at risk.

The dramatic increases have raised concerns about the sustainability of current systems. Differential rates of increase across forms of maltreatment have raised questions about whether traditional models of service delivery are appropriate to meet the changing needs and deal with the changing risks documented in child maltreatment cases. In Alberta and Ontario, CIS information has been instrumental in recent shifts from a "one size fits all" model to differential response models that consider the diverse and multiple needs of children and families. In addition, despite the increase in rates of substantiation for exposure to domestic violence, the CIS documents that relatively small proportions of these cases receive ongoing child welfare service. These findings have led to changes in the ways that exposure to domestic violence cases are handled at the investigation stage (e.g., in Ontario). Thus, documenting shifts in the profiles of maltreatment and clients has served as an impetus for exploring new approaches to service delivery.

CIS data have also been used extensively in secondary data analysis to identify determinants of service decisions (Fallon, 2005), case characteristics and disparate service responses for Aboriginal compared to non-Aboriginal children and families (Blackstock, Trocmé, & Bennett, 2004; Trocmé, Knoke, & Blackstock, 2004), patterns of maltreatment for children with developmental delays (Fudge-Schormans & Brown, 2002), characteristics of children placed in out-of-home care and their families (MacLaurin, Fallon, & Trocmé, 2003), the incidence of investigations for cases involving physical punishment (Trocmé & Durrant, 2003), and to estimate the rate and severity of physical harm resulting from maltreatment (Trocmé, MacMillan, Fallon, & De Marco, 2003). These studies and future analyses based on CIS data are expected to influence policy and service delivery.

Given the function that the CIS serves in the Canadian child welfare sector, it is important to ensure that the data quality is sufficiently high to meet the needs of researchers and policy makers. Evaluation of the psychometric properties of the CIS instrument is essential to establish confidence in the data collected and the estimates generated. The present study examines the reliability of the CIS data collection instrument. Confidence in the estimates derived is based upon the supposition that case information recorded by workers at the time of investigation reflects real differences in case characteristics rather than measurement error. Low levels of reliability reduce the precision of the estimates generated. Adequate reliability is a necessary but not a sufficient condition for establishing the validity of an instrument. The validity of the information collected is contingent on the amount of measurement error present in the data collected.

Two forms of reliability may be used to quantify the measurement variance in CIS data: inter-rater reliability and test re-test reliability. Inter-rater reliability provides an index of the consistency of ratings across workers at a single point in time. Assessment of inter-rater reliability is precluded by the context in which the CIS forms are completed. The CIS collects information from individual workers regarding the characteristics of cases to which they are assigned in practice, upon completion of the child maltreatment investigation. Assessment of inter-rater reliability would be possible only if two workers had similar levels of case involvement and knowledge. Test re-test reliability provides an index of response stability over time and can be examined by having the same worker apply CIS criteria to evaluate a case at two points in time. If the CIS definitions and criteria are clear, workers are expected to display consistency in the ways they classify their clients' characteristics and report on their service decisions at the point of investigation. The present study examines the results of test re-test reliability assessment during pilot implementation of the CIS Cycle II instrument.

METHODS

The CIS data collection instrument gathered information on 130 variables, grouped into four primary categories: (a) characteristics of the alleged maltreatment, (b) case disposition variables, (c) characteristics of families and caregivers, and (d) characteristics of the child victims of the alleged maltreatment. The reliability of the Cycle II instrument was assessed as part of a larger study that included focus testing of two versions of the instrument in two sites. The majority of items on the two versions were identical,[1] yielding a sample of 82 children from 57 households for most items. When one version is examined, analyses are performed on the subset of cases. For these items the version and corresponding sample size are noted in the tables. Version A included 54 children from 34 families, and Version B assessed 28 children from 23 families.

A convenience sample of three metropolitan child welfare agencies was selected for reliability testing based upon availability and proximity to the study team. Workers in these three sites were then asked to participate voluntarily. Participating workers completed the CIS data collection instrument on a total of 57 families who were the subject of a new intake investigation for suspected or alleged child maltreatment. The CIS instrument was designed to be completed on all children in the family who were at risk for child maltreatment. Thirty percent of workers completed forms for more than one child in a family. Following completion of the CIS instrument, the completed CIS instruments were submitted to the CIS study team. Participating workers were then asked to complete a second CIS instrument on the same 57 families approximately 3 weeks after the first completed report. The second CIS instrument was completed 4.5 weeks on average following the first completed report. Because the primary objective of the CIS is to have workers record—systematically and using standardized criteria—the information collected during investigations, workers utilized case information available in files and/or administrative databases. In other words, the primary objective of the study was not to assess whether workers accurately recalled case information but whether information that workers collect as part of an investigation could be reported consistently according to CIS criteria.

Indices of Agreement

Because child age was recorded as a continuous variable, Pearson's correlation was used as the index of test re-test reliability for that variable. The majority of variables were nominal or ordinal in nature, requiring workers to select among categories. Cohen's Kappa (κ) statistic was computed for nominal variables using the Statistical Package for the Social Sciences (SPSS). In adjusting for chance agreement, the κ statistic provides a more rigorous measure of the concordance between two sets of ratings than percent agreement. That is, level of agreement that occurs by chance is given a value of 0 and the value of κ indicates how much the attained level of agreement differs from chance. Linear weighted κ was calculated for ordered variables with three or more levels, using the Vassarstats κ calculator. The weighted κ takes into account the ordinal nature of variables, penalizing less for misclassifications in adjacent categories than for ratings that differ by two or more levels. Cases for which "unknown" was endorsed were omitted from the analyses (family income, number of family moves, number of previous case file openings, length of time since last opening, duration of primary maltreatment). Kappa values were classified according to Landis and Koch (1977) as excellent (0.81–1.00), substantial (0.61–0.8) and moderate agreement (0.41–0.60). Values of less than 0.40 are considered to reflect poor agreement.

The reliability of some CIS data elements could not be assessed. The computation of κ requires a symmetric table with two or more levels, in which response options endorsed at Time 1 match the number of response options endorsed at Time 2. Analyses of five items yielded two by one tables, precluding the calculation of κ.[2] In addition, seven characteristics were not noted for any cases in this sample, at Time 1 or Time 2.[3] Although consistent judgements regarding the absence of particular characteristics demonstrate consistency, CIS items that could not be assessed using κ are excluded from the tables and from the CIS variable descriptions below. In some cases, the reliability of higher-order categories of items was assessed, rather than or in addition to assessment of individual items. For example, specific acts of maltreatment were examined, and these acts were grouped into four broad categories or forms of maltreatment. Workers' consistency in identifying an act in the same maltreatment category was examined, in addition to their consistency in reporting the specific act of maltreatment. These additional analyses were applied to a select subset of items, to assess the reliability of approaches that have been used in research using CIS data. In total, test re-test reliability was assessed for 120 variables. These variables and the response options provided for each are described below.

Maltreatment Characteristics

Form of Maltreatment

Workers reported the primary type of maltreatment alleged for each child and up to two secondary types. Workers were asked to record the code that corresponded to the type of maltreatment alleged/suspected, with primary type placed first. Each form listed 21 specific types of maltreatment, each subsumed within one of four broad forms of maltreatment: physical abuse, sexual abuse, neglect, or emotional maltreatment. Consistency of worker judgements regarding the nature of maltreatment was examined three ways. First, the primary form of maltreatment was assessed. Second, the extent of agreement regarding whether any neglect, physical abuse, sexual abuse, or emotional maltreatment was suspected or alleged, as a primary *or* secondary form, was examined. Third, test re-test reliability was examined for 15 of the 16 specific types of maltreatment identified in this sample of cases: shaken baby syndrome, abusive physical punishment, other physical abuse, touching/fondling of genitals, sexual exploitation, failure to supervise/protect from sexual abuse, physical neglect, medical neglect, permitting criminal behaviour, abandonment, educational neglect, failure to thrive, emotional neglect, emotional abuse, and exposure to domestic violence.

Two additional aspects of the alleged/suspected maltreatment were examined for the primary form of maltreatment: the estimated duration of primary maltreatment (single incident, less than 6 months, more than six months, unknown, or not applicable, if maltreatment was unfounded), and whether unfounded investigations were malicious referrals (yes, no, or unknown).

Physical Harm

If the child sustained physical harm, workers indicated the type of injury experienced (bruises/cuts/scrapes, broken bones, head trauma, or other health conditions). "Health or safety seriously endangered" was recorded as yes, no, or not applicable, no harm.

Emotional Harm

Symptoms of mental or emotional harm were noted using three sequential items reflecting increasing severity of emotional distress: "no current signs, but mental or emotional harm is probable"; "child shows signs of mental or emotional harm"; and "exhibited mental or emotional harm, requires therapeutic treatment."

Perpetrator Identity

Workers specified the alleged perpetrator(s) by checking Caregiver A, Caregiver B, and/or Other Alleged Perpetrator(s).

Sources of Referral for Alleged Maltreatment

Workers selected from a list of 18 referral sources that were grouped into seven categories for analyses: family (custodial parent, non-custodial parent, or child), relative/neighbour/friend, health/mental health professional (public health nurse, physician, mental health professional), other professional or community referral (community/recreation centre, community agency, other child welfare service, crisis service/shelter, social assistance worker), school/day care centre, police, and other.

Case Disposition

Seven case decisions were included in this category: (a) investigated maltreatment was classified as "substantiated" if the balance of evidence indicated that it occurred, "suspected" if the evidence was insufficient for the worker to substantiate but the occurrence of maltreatment could not be ruled out, or "unfounded" if the balance of evidence indicated that maltreatment had not occurred; (b) workers indicated if the case was to be kept open for service at the completion of the investigation; (c) decisions regarding out-of-home placement were recorded as "no," "considered," or "required"; (d) if out-of-home placement was required, the placement type was selected (informal kinship care, formal kinship care, other family foster care, group home, residential/secure treatment); (e) application to child welfare court was recorded as "no court considered," "application considered," "application made," or "mediation/alternate response"; (f) workers indicated whether there was a police investigation regarding child maltreatment (yes or no); and (g) "police charges laid" included yes, no, and not applicable response options.

Family and Caregiver Variables

Household Characteristics

Workers provided information regarding select family and household characteristics. Socioeconomic status was assessed by estimated family income. Four annual income categories were provided, ranging from "less than $15,000" to "greater than $40,000" in ascending order. An "unknown" income category was also included. Adequacy of housing was assessed using several variables: type of housing accommodation (own home, rental housing, public housing, shelter/hotel, other housing, or unknown); unsafe housing conditions (yes, no, or unknown); home overcrowding (yes, no, or unknown); and number of moves for a family in the previous year (never, once, twice, three or more moves, or unknown).

Several additional family characteristics were noted. Each caregiver absent from the family home was noted separately (mother, father, other). Workers noted whether other adults (e.g., grandparent, boarder) resided in the home by checking a box if affirmative. Response options for "ongoing child custody dispute" and "spanking is employed as a form of discipline within the home" were yes, no, and unknown.

Caregiver Sociodemographic Characteristics

Sociodemographic information was collected for up to two caregivers residing with the index child(ren). Only the test re-test reliability for the primary caregiver was examined in the present study. Caregiver sex was noted as male or female. Relationship to the child included seven options: biological parent, adoptive parent, stepparent, common-law partner, foster parent, grandparent, and other. Based upon the Statistics Canada 1996 census, eight ethno-racial categories and "other" were provided. English, French, and "other" were included as primary language options. Sources of caregiver income were full-time, part-time and seasonal employment, social assistance, unemployment insurance, other benefit, no income, and unknown. Test re-test reliability for age category and educational attainment of the primary caregiver was assessed using the subsample that completed Version B. Ten caregiver age categories were provided: Under 16 years of age, 16 to 18, 19 to 21, 22 to 25, 26 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, and older than 70 years of age. Categories of educational attainment were elementary or less, secondary or less, college or university, and unknown.

Caregiver Functioning

Nine caregiver functioning concerns were examined: alcohol abuse, drug abuse, criminal activity, cognitive impairment, mental health concerns, physical health concerns, few social supports, caregiver maltreated as child, and primary caregiver in violent relationship, with response options being no, suspected, confirmed, and unknown for each concern. Caregiver functioning concerns were categorized as confirmed if the concern was (a) diagnosed, (b) disclosed, (c) observed by the worker or another worker, or (d) on the file. Concerns were categorized as suspected if the evidentiary criteria for confirmation could not be met but, at the conclusion of the investigation, workers thought that a particular concern was likely. In addition, the nature of the contact with the caregiver was rated as cooperative, not cooperative, or not contacted.

Family Maltreatment History

Family history of child maltreatment was represented by three variables: (a) the number of times a case was previously opened (never, once, 2–3 times, more than three times, or unknown); (b) the number of months since the last case opening (less than three months, 3–6 months, 7–12 months, 13–24 months, or more than 24 months); and (c) whether previous maltreatment was substantiated for any child in the family (yes, no, unknown, or "previous opening was not a maltreatment investigation"). In addition, workers indicated whether domestic violence was investigated and whether police charges for domestic violence were laid, with yes or no response options for each.

Service Referrals

Workers checked a box to indicate the services to which families and children were referred, if any. Family-focused referrals included parent support group, in-home parenting support, other family or parent counselling, alcohol/drug counselling, welfare/social assistance, food bank, shelter services, and domestic violence services. Child-focused referrals included psychiatric/psychological service, special education placement, victim support program, medical/dental services, and other child counselling. Reliability was assessed for each service referral, for any referral, for any family-focused referral, and for any child-focused referral.

Child Variables

Age and Sex

Age and sex were recorded for each child.

Child Functioning

For each child, workers indicated whether each of 20 child functioning concerns was suspected, confirmed, not present, or unknown. These concerns were developmental delay, physical disability, learning disability, substance-abuse-related birth defect, other health condition, specialized education services, depression or anxiety, self-harming behaviour, psychiatric disorder, positive toxicology at birth, negative peer involvement, attention deficit disorder with or without hyperactivity, drug/solvent abuse, violence toward others, running away, irregular school attendance, inappropriate sexual behaviour, *Youth*

Criminal Justice Act involvement, and other behavioural or emotional problems.

Maltreatment History for Index Child

Maltreatment history for each investigated child was represented by (a) a previous report of maltreatment for the child, and (b) whether previous maltreatment for that child was substantiated (yes, no, or unknown).

RESULTS

The levels of agreement for maltreatment characteristics are presented in Tables 1 and 2. With the exception of the identification of Caregiver A as the perpetrator of maltreatment, the maltreatment variables included in Table 1 are characterized by substantial to excellent agreement. Excellent test re-test reliability was evident in workers' classification of the primary form of maltreatment under investigation and the estimated chronicity of that maltreatment. Workers showed substantial to excellent agreement regarding whether any neglect or any physical, sexual, or emotional maltreatment was suspected or alleged, as a primary or secondary form and regarding cases involving malicious referrals. With the exception of "other referral," the reliability of information about the source of the referral to child welfare services was excellent.

Judgements about the presence and nature of physical harm and the likelihood of emotional harm were generally consistent. However, less severe physical injuries (i.e., bruises, cuts, or scrapes) were less likely than more severe injuries to be noted consistently. Judgements about the involvement of Caregiver A in maltreatment were less consistent than those made about Caregiver B and the "other perpetrator" category. Inspection of the data indicates that for 11 of the 12 cases with inconsistent ratings for Caregiver A, the caregiver was included as one of two perpetrators at Time 1. However, responses at time 2 indicated that Caregiver A was not a suspected or alleged perpetrator.

Types of maltreatment are presented in Table 2, by the level of agreement attained. Twelve of the 16 types examined were characterized by substantial to excellent test re-test reliability. Moderate test re-test reliability was evident for "other physical abuse," "sexual exploitation," and "failure to protect: physical." Only "emotional neglect" showed poor agreement, at chance levels. Examination of data indicated that fewer than one third of cases involving "emotional neglect" were assigned the same type of primary maltreatment on both occasions.

Table 1
Test Re-test Reliability of Maltreatment Variables

Variable	Kappa
Primary form of maltreatment (neglect or physical, sexual, or emotional maltreatment)	0.88
Form of maltreatment (primary or secondary)	
Any physical abuse	0.92
Any sexual abuse	1.00
Any neglect	0.61
Any emotional maltreatment	0.67
Duration of primary maltreatment (if known and maltreatment was suspected or substantiated; N = 43)	0.89[a]
If unfounded, malicious referral? (N = 22)	0.91
Physical harm/endangerment	
No physical harm	0.72
Bruises, cuts, scrapes	0.63
Broken bones	1.00
Head injury	1.00
Health or safety physically endangered[b]	1.00
Emotional harm	
Signs of emotional harm	0.76
Emotional harm probable	0.74
Treatment required for emotional harm	0.73
Perpetrator identity	
Caregiver A	0.52
Caregiver B	0.61
Other perpetrator	0.70
Source of referral to Child Welfare Services	
Family	1.00
Relative/neighbour/friend	0.91
Health/mental health professional	1.00
Other professional or community	0.82
School/day care centre	0.95
Police	1.00
Other	0.66

[a]linear weighted kappa (κ). [b] Version B used, N = 28.

Table 2
Types of Maltreatment by Level of Reliability

Excellent agreement (≥ 0.8)	Substantial agreement (0.6–0.79)	Moderate agreement (0.4–0.59)	Poor agreement (< 0.4)
Shaken baby syndrome (1.00)	Permitting criminal behaviour (0.79)	Failure to protect: physical (0.54)	Emotional neglect (0.38)
Medical neglect (1.00)	Touching/fondling genitals (0.74)	Sexual exploitation (0.49)	
Failure to protect: sexual (1.00)	Abandonment (0.66)	Other physical abuse (0.48)	
Educational neglect (1.00)			
Failure to thrive (1.00)			
Emotional abuse (0.90)			
Exposure to family violence (0.86)			
Physical neglect (0.82)			
Abusive physical punishment (0.80)			

Table 3 presents the κ values for case disposition variables. Excellent test re-test reliability (≥ 0.8) was evident for judgements about the decision to open a case for ongoing service, whether out-of-home placement was required and, if required, the type of placement provided. Substantial agreement (0.6–0.79) was attained for the level of substantiation for primary maltreatment, application to child welfare court, for police investigation and police charges for child maltreatment.

Table 3
Test Re-test Reliability of Case Disposition Variables

Variable	Kappa
If placed, type of out-of-home placement (*N* = 19)	1.00
Out-of-home placement required	0.97
Case to stay open for ongoing child welfare services	0.90
Substantiation of primary maltreatment	0.80
Police charges laid for child maltreatment	0.79
Police investigation for child maltreatment	0.77
Application to Child Welfare Court	0.76

Family-level factors are presented in summary form by level of reliability in Table 4. Among the factors assessed, substantial to excellent test re-test reliability was evident for 91% (30 of 33) of the items. Three items were classified as moderate in their level of agreement: unsafe housing, home overcrowded, and involvement of the primary caregiver in criminal activity. None of these factors was found to have poor reliability (< 0.4).

Table 4
Family-Level Factors by Level of Reliability

	Excellent agreement (> 0.8)	Substantial agreement (0.6–0.8)	Moderate agreement (0.4–0.59)
Household characteristics	Adult domestic violence investigation (1.00)[b]	Other adults in the home (0.78)	Unsafe housing conditions (0.57)
	Police charges for adult domestic violence (1.00)[b]	Spanking used as a form of discipline (0.78)	Home overcrowded (0.55)
	Ongoing custody dispute (0.89)	Number of moves in past 12 months (0.74)[a,b]	
	Type of housing (0.86)	Other caregiver outside of the home (0.70)	
	Mother outside of home (0.85)		
	Father outside of home (0.85)		
	Estimated family income (0.84)[a]		
Primary caregiver socio-demographic characteristics	Relationship to child (1.00)	Primary language (0.70)	
	Age (0.95)[c]	Educational attainment (0.69)[c]	
	Sex (0.91)		
	Ethno-racial group (0.91)		
	Primary income source (0.80)		
Primary caregiver functioning	Mental health concerns (1.00)[c]	Physical health concerns (0.76)[c]	Criminal activity (0.57)[c]
	Drug abuse (0.91)[c]		
	Cognitive impairment (0.88)[c]	Alcohol abuse (0.71)[c]	
	Caregiver maltreated as child (0.88)[c]	Caregiver in a violent relationship (0.60)[c]	
	Caregiver cooperation (0.84)		
	Few social supports (0.80)[c]		
Family maltreatment history	Number of prior case openings (0.85)[a]	Substantiation of prior maltreatment (0.73)	
	Length of time since last case opening (0.80)[a]		

[a] linear weighted Kappa. [b] Version A used, *N* = 34. [c] Version B used, *N* = 23.

The level of reliability for each type of service referral is examined in Table 5. Substantial or excellent reliability was exhibited on 62% of the individual service referrals. This means that approximately one in three judgements about supportive services was inconsistent. When the individual service referrals were grouped into the categories "any referral," "family-focused referral," or "child-focused referral," the reliability of the first two categories were substantial and excellent, respectively. Thus, workers may accurately record that a referral was made, but the specific nature of the referral was classified differently at each point in time. Agreement about whether child-focused referrals were made was moderate.

Table 5
Type of Service Referral by Level of Reliability

	Excellent agreement (> 0.8)	Substantial agreement (0.6–0.8)	Moderate agreement (0.4–0.59)	Poor agreement (< 0.40)
Service referrals				
Family-focused	Other family/ parenting counselling (0.82)	Drug/alcohol counseling (0.79) Welfare/social assistance (0.65) Domestic violence services (0.63)	In-home parenting support (0.56)	Shelter services (0.37) Parent support group (0.35)
Child-focused		Medical/dental services (0.74) Special education placement (0.66) Other child counseling (0.66)		Psychological services (0.24) Victim support program (0.23)
Referral category	Any family-focused referral (0.93)	Any referral for services (0.80)	Any child-focused referral (0.48)	

Table 6 presents levels of agreement for child variables. Twenty of the 24 child-level factors had substantial to excellent test re-test reliability, with the majority falling within the substantial range. Four child functioning concerns were moderate in level of agreement, indicating a larger margin of error for these than for other child behavioural concerns.

Table 6
Child-Level Factors by Level of Reliability

	Excellent agreement (> 0.8)	Substantial agreement (0.6–0.8)	Moderate agreement (0.4–0.59)
Child characteristics	Sex (0.93) Age (r = 0.996)		
Child functioning concerns[a]	Physical disability (0.81)	Age-inappropriate sexual behaviour (0.76) Special education services (0.76) Depression/anxiety (0.73) Irregular school attendance (0.72) Other health conditions (0.72) Developmental delay (0.71) Violence toward others (0.68) Other behaviour problems (0.67) Substance-abuse-related birth defects (0.67) Negative peer involvement (0.63) Positive toxicology at birth (0.61) Learning disability (0.61) Psychiatric disorder (0.61) ADD/ADHD (0.60)	Self-harm behaviour (0.55) Substance abuse (0.51) Running away (0.51) *Youth Criminal Justice Act* involvement (0.46)
Child maltreatment history	Substantiation of prior maltreatment (0.89)[a] Child previously reported (0.85)[a]		
Other child maltreatment variables	Physician physically examined child (0.96)		

[a] Version B used, N = 28.

DISCUSSION

The present study was undertaken to examine the test re-test reliability of data collected using the CIS Cycle II instrument. Consistency of worker reporting was evaluated by comparing case information provided at two independent points in time. Test re-test reliability was examined for select case disposition variables, maltreatment history, and a variety of characteristics of the suspected/alleged maltreatment, households, caregivers, and children.

The vast majority of items on the CIS Cycle II form showed substantial to excellent test re-test reliability. The form and estimated duration of primary maltreatment, indices of physical and emotional harm, referral source, case disposition variables, and the majority of family, primary caregiver, and child characteristics were rated consistently over time. Substantial to excellent agreement was also shown in 75% of the individual types of maltreatment and 62% of the service referrals.

Few variables were characterized by poor reliability. Examined individually, several services to which families and children were referred and some specific acts of maltreatment were rated inconsistently over time. Caution must be exercised in using these variables individually. Where reliability was examined for higher-order categories, it was frequently enhanced by combining items. For example, although some specific acts of maltreatment were not recorded consistently by workers, the category of maltreatment was rated consistently over time (e.g., sexual abuse or physical abuse). Similarly, service referrals grouped into "any referral" or "any family referral" showed higher levels of agreement than judgements about some of the specific service referral types. Level of agreement regarding unsafe and overcrowded housing, criminal activity of the primary caregiver, any child-focused referral, perpetrator identity (Caregiver A), and several child functioning concerns fell within the moderate level of agreement, below the criterion adopted for acceptable reliability in this study.

In interpreting these findings, several aspects of the study design and their potential impact on reliability estimates warrant consideration. First, given the small sample size and the low frequency with which some of the items are noted, a small number of misclassified cases have a substantial impact on estimates of reliability. Second, the average length of time between ratings was 4.5 weeks. Given this interval between ratings, it is difficult to ascertain whether inconsistencies were related to information decay, the availability of new information, or changes in the status of certain case variables in the interval between worker ratings. An additional limitation of this study is the use of a convenience sample of workers. There may be differences in reliability across workers; thus, reliability estimates based upon this subset of volunteers are not generalizable. For example, by virtue of their consent to participate in the study, workers may have attended differently to information collected during their investigations.

Test re-test reliability is appropriate if the phenomenon that is being measured is stable in the interval between assessments. Broedling

(2001) notes the importance of recognizing that test re-test reliability estimates include two potential sources of variation: measurement variance and true variance. Measurement variance is related to factors such as the inconsistent application of rating criteria by the same worker or errors in recall for relevant information. True variance, on the other hand, is related to changes in the true scores of clients over time. For example, in the context of child welfare practice, case characteristics may change as a function of intervention (e.g., referrals) or because further assessment provides additional information. In the present study the workers who completed the CIS forms had no further service contact with families after the completion of investigation. However, in completing the CIS instrument at the second point in time, workers may have had access to post-investigation case information through interaction with colleagues or paper or electronic files. Changes in the case information over time may have influenced workers' assessments at different intervals. A shorter interval may reduce the possibility that differences in ratings reflect changes in characteristics of families and children.

The CIS samples a wide variety of family, caregiver, child, and maltreatment characteristics. There are few other studies of comparable instruments against which to compare the present findings. Comparisons with similar instruments would serve two purposes. First, it would assist in identifying whether particular constructs, in general, tend to be associated with greater measurement error. Second, it would provide an index of the range of reliabilities found in studies of other instruments used for similar purposes. Risk assessment instruments in child welfare provide worker assessments of similar constructs and, in theory, may provide a basis for comparison. However, there has been limited research on the psychometric properties of many tools commonly used to assess risk for child maltreatment. Available studies of reliability focus on evaluations of internal consistency and inter-rater reliability, and most examine the reliability of overall risk ratings rather than the reliability of individual items.

Test re-test reliability of a variety of self-report instruments has been evaluated, providing some index of the range of reliability estimates typically attained. For example, instruments measuring adolescent risk behaviours (Brener, Collins, Kann, Warren, & Williams, 1995; Brener et al., 2002; Flisher, Evans, Muller, & Lombard, 2004), adolescent self-reported substance use (e.g., O'Malley, Bachman, & Johnston, 1983), adolescent subjective psychological and physical health complaints (Haugland & Wold, 2001), and surveys of health behaviours and use of counselling and health services (Santelli, Klein, Graff, Allan, & Elster, 2002) have been found to have test re-test reliabilities ranging from the moderate (κ values exceeding 0.4) to the "almost perfect" range for most individual items. Similarly, the test re-test reliability of adult self-reported health status ranged from 0.58 for "frequent mental distress" to 0.75 for "self-reported health" and "healthy days" (Andreson, Catlin, Wyrwich, & Jackson-Thompson, 2003). A telephone survey of self-reported health risks found substantial to almost perfect agreement for responses to questions about demographic factors and health behaviours (Starr, Dal Grande, Taylos, & Wilson, 1999). Using a variety of methods, the test-retest reliability of questions pertaining to substance use, abuse, and dependence is generally good to excellent (e.g., Aktan, Calkins, Ribisl, Kroliczak, & Kasim, 1997). These instruments include a relatively small number of items relative to the CIS instrument and require participants to report on their own behaviours. In contrast, workers completing the CIS instrument use a combination of sources of information to inform their ratings of clients. Despite these differences, these studies suggest that the reliability estimates documented for the CIS fall within the range of those documented for a variety of other assessment instruments.

A fundamental question that emerges from these analyses is whether the reliability estimates attained in this study are sufficient, given the widespread use of CIS data in policy and research contexts. In general, items used to generate national incidence rates showed excellent concordance (primary form of maltreatment, maltreatment substantiation, case openings for ongoing service, out-of-home placement, prior child welfare reports, source of referral, and estimated duration of maltreatment). These findings indicate that workers were able to consistently report these features of their investigations. Many family and child characteristics were also found to have excellent reliability, suggesting that the CIS data collection instrument has utility in profiling the families and children investigated by child welfare authorities. Reliability estimates for physical and emotional harm were slightly less reliable, suggesting that incidence rates for these items will have a larger margin of error. In addition, some items that required a greater specificity, such as the type of service referral or the precise nature of child functioning problems, were less likely to be reported consistently (i.e., values in the poor range or at the low end of substantial). Caution must be exercised in the use of these items in secondary analyses.

Provincial and federal policy makers across Canada rely on data from the CIS as a key source of information on the changing profile of children and families served by child welfare authorities. In the absence of better integrated information systems, surveys such as the CIS that make use of data collected for administrative purposes can provide a wealth of information to support policy makers and agency administrators. Reliability studies are necessary to demonstrate data integrity and guide researchers in selecting variables for analysis.

ACKNOWLEDGEMENTS

We gratefully acknowledge support for the 2003 Canadian Incidence Study of Reported Child Abuse and Neglect from the Public Health Agency of Canada, Injury and Child Maltreatment Section, Health Surveillance and Epidemiology Division, and support for secondary data analyses from the Canadian Institutes of Health Research under grant 43277.

NOTES

1. One version was administered in Ontario (Version A) and the other in Alberta (Version B). Eighteen of the 130 CIS data elements (14%) contained slightly different response options on the two versions. For items with different response options on Versions A and B, test re-test reliability was examined for the version containing the format adopted on the final CIS Cycle-II form.

2. These items were a history of undetected or misdiagnosed injuries, medical treatment required for injury, family referral to a food bank, and two types of sexual abuse (sexual activity completed and sexual harassment).

3. The following case characteristics were not noted for any of the cases in this sample: whether the child experienced burns/scald or fatal abuse, referrals from anonymous sources, maltreatment involving attempted sexual activity, voyeurism/exhibitionism, failure to provide psychological treatment, and child referrals to recreational services.

REFERENCES

Aktan, G.B., Calkins, R.F., Ribisl, K.M., Kroliczak, A., & Kasim, R.M. (1997). Test-retest reliability of psychoactive substance abuse and dependence diagnoses in telephone interviews using a modified Diagnostic

Interview Schedule-Substance Abuse Module. *American Journal of Drug & Alcohol Abuse, 23*(2), 229–248.

Alberta Children's Services Annual Report. 2000–2001. Retrieved April 28, 2008, from <http://www.child.alberta.ca/home/documents/rpt_00-01_annual.pdf>

Alberta Children's Services Annual Report. 2006-2007. Retrieved April 28, 2008, from <http://www.child.alberta.ca/home/documents/rpt_06-07_annual.pdf>

Andreson, E.M., Catlin, T.K. Wyrwich, K.W., & Jackson-Thompson, J. (2003). Retest reliability of surveillance questions on health related quality of life. *Journal of Epidemiology and Community Health, 57*(5), 339–343.

Blackstock, C., Trocmé, N., & Bennett, M. (2004). Comparison between Aboriginal, visible minority and non visible minority children in CIS-98. *Violence Against Women, 10*(8), 901–916.

Brener, N.D., Collins, J.L., Kann, L., Warren, C.W., & Williams, B.I. (1995). Reliability of the Youth Risk Behaviour Questionnaire. *American Journal of Epidemiology, 141*, 575–580.

Brener, N.D., Kann, L., McManus, T., Kichen, S.A., Sundberg, E.C., & Ross, J.G. (2002). Reliability of the 1999 Youth Risk Behavior Survey. *Journal of Adolescent Health, 31*, 336–342.

Broedling, L.A. (2001). On more reliably employing the concept of "reliability." *Public Opinion Quarterly, 38*(3), 372–378.

English, D.J., Marshall, D.B., Brummel, S., & Orme, M. (1999). Characteristics of repeated referrals to Child Protective Services in Washington State. *Child Maltreatment, 4*(4), 297–307.

Fallon, B. (2005). Factors driving case dispositions in child welfare services: Challenging conventional wisdom about the importance of workers and organizations. *Dissertation Abstracts International, 66*(06), 2384A.

Flisher, A.J., Evans, J., Muller, M., & Lombard, C. (2004). Brief report: Test-retest reliability of self-reported adolescent risk behaviour. *Journal of Adolescence, 27*, 207–212.

Fluke, J.D., Yuan, Y.T, & Edwards, M. (1999). Recurrence of maltreatment: An application of the National Child Abuse and Neglect Data System (NCANDS). *Child Abuse & Neglect, 23*, 633–650.

Flynn, R., & Bouchard, D. (2005). Randomized and quasi-experimental evaluations of program impact in child welfare in Canada: A review. *Canadian Journal of Program Evaluation, 20*(3), 65–100.

Fudge-Schormans, A., & Brown, I. (2002). Maltreatment in children with developmental delays. *Journal on Developmental Disabilities, 9*(1), 1–19.

Haugland, S., & Wold, B. (2001). Subjective health complaints in adolescents: Reliability and validity of survey methods. *Journal of Adolescence, 24*(5), 611–624.

Iron, K., Przybysz, R., & Laupacis, A. (2003). *Access to MRI in Ontario: Addressing the information gap*. Institute for Clinical Evaluative Services. Retrieved September 15, 2007, from <http://www.ices.on.ca/file/Access%20to%20MRI%20in%20Ontario%20-%20Addressing%20the%20information%20gap_printer%20friendly.pdf>

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

MacLaurin, B., Fallon, B., & Trocmé, N. (2003). Characteristics of investigated children and families referred for out-of-home placement. In K. Kufeldt & B. McKenzie (Eds.), *Child welfare: Connecting research, policy and practice* (pp. 27–40). Waterloo, ON: Wilfrid Laurier University Press.

Nguyen, T.D., Attkinson, C.C., & Bottino, M.J. (1983). The definition and identification of human service needs in a community. In R.A. Bell, M. Sundel, J.F. Aponte, S.A. Murrell, & E. Lin (Eds.), *Assessing health and human service needs: Concept, methods and applications* (pp. 88–110). New York: Human Sciences Press.

O'Malley, P.M., Bachman, J.G., & Johnston, L.D. (1983). Reliability and consistency in self-reports of drug use. *International Journal of the Addictions, 18*, 805–824.

Ontario Ministry of Children and Youth Services. (2005). *Child Welfare Transformation 2005: A strategic plan for a flexible, sustainable and* *outcome oriented service delivery model*. Retrieved January 5, 2007, from <http://govdocs.ourontario.ca/results?fsu=Child+welfare>

Pecora, P.J. (2002, September). *The challenge of achieving evidence-based practice in child and family services*. Paper presented at ACWA 2002 conference: What Works!? Evidence-Based Practice in Child and Family Services, Sydney, Australia.

Santelli, J., Klein, J., Graff, C., Allan, M., & Elster, A. (2002). Reliability in adolescent reporting of clinician counseling, health care use, and health behaviors. *Medical Care, 40*(1), 26–37.

Sieppert, J. (2005), Evaluation in Canada's social services: Progress, rifts, and challenges. *Canadian Journal of Program Evaluation, 20*(3), 101–121.

Starr, G.J., Dal Grande, E., Taylos, A.W., & Wilson, D.H. (1999). Reliability of self reported behavioural health risk factors in a South Australian telephone survey. *Australian and New Zealand Journal of Public Health, 23*(5), 528–530.

Thompson, J., & Fallon, B. (1999). *The first Canadian roundtable on child welfare outcomes: Roundtable proceedings*. Toronto: Bell Canada Child Welfare Research Unit.

Trocmé, N., & Durrant, J. (2003). Physical punishment and the response of the Canadian child welfare system: Implications for legislative reform. *Journal of Social Welfare and Family Law, 25*(1), 1–18.

Trocmé, N., Fallon, B., MacLaurin, B., Daciuk, J., Felstiner, C., Black, T., et al. (2005). *Canadian Incidence Study of Reported Child Abuse and Neglect, CIS-2003: Major findings report*. Ottawa, ON: Public Health Agency of Canada, Government of Canada.

Trocmé, N., Knoke, D., & Blackstock, C. (2004). Pathways to the overrepresentation of Aboriginal children in Canada's child welfare system. *Social Service Review, 78*(4), 577–601.

Trocmé, N., MacLaurin, B., & Fallon, B. (2000). Canadian Child Welfare Outcomes Indicator Matrix: An ecological approach to tracking service outcomes. *Journal of Aggression, Maltreatment and Trauma, 4*(1), 165–190.

Trocmé, N., MacLaurin, B., Fallon, B., Daciuk, J., Billingsley, D., Tourigny, M., et al. (2001). *Canadian Incidence Study of Reported Child Abuse and Neglect: Final report*. Ottawa, ON: Minister of Public Works and Government Services Canada.

Trocmé, N., MacMillan, H., Fallon, B., & De Marco, R. (2003). Nature and severity of physical harm caused by child abuse and neglect: Results from the Canadian Incidence Study. *Canadian Medical Association Journal, 169*(9), 911–915.

Vandermeulen, G., Wekerle, C., & Ylagan, C. (2005). Introduction to the special issue on child welfare-research collaborations: Teamwork, research excellence, and credible, relevant results for practice. *Ontario Association of Children's Aid Societies (OACAS) Journal, 49*(1), 2–3.

Della Knoke, M.A., Ph.D., is a public servant for the Government of Ontario. This work was completed when she was a Research Associate at the Centre of Excellence for Child Welfare. Her current work focuses on policy and program evaluation and conducting research for policy development.

Nico Trocmé, M.S.W., Ph.D., is the Philip Fisher Chair in Social Work at McGill University, the director of the McGill Center for Research on Children and Families, and the scientific director of the Centre of Excellence for Child Welfare. His research focuses on the epidemiology of child maltreatment and the evaluation of child welfare policies and services in Canada.

Bruce MacLaurin, M.S.W., is a Ph.D. candidate and an assistant professor in the University of Calgary's Faculty of Social Work. His research interests include child maltreatment, child welfare policy and service delivery, foster care outcomes, street youth, and youth at risk.

Barbara Fallon, M.S.W., Ph.D., is an assistant professor at the University of Toronto's Faculty of Social Work and the Director of the Canadian Incidence Study of Reported Child Abuse and Neglect. Her research interests focus on program evaluation and organizational research in child welfare.