Studying Factors Affecting Creation and Fate of Innovations and their Organizations–

II: Verification of Raters and the Instrument

Eleanor D. Glor

Editor-in-Chief
The Innovation Journal: The Public Sector Innovation Journal

Studying Factors Affecting Creation and Fate of Innovations and their Organizations – II: Verification of Raters and the Instrument[1] [2]

Eleanor D. Glor

ABSTRACT

This paper evaluates a new quantitative measurement instrument measuring factors influencing introduction, implementation and fate of innovations and their organizations (Glor, 2017). The instrument measures six types of factors: ideology, politics, support, economic and fiscal situation, program and organizational resources, and program and organizational effects. The instrument is evaluated by having three expert raters complete it for five case studies, the sub-population of income security innovations and their organizations introduced by the Government of Saskatchewan (GoS), Canada, 1971-82. The verification considers *rater* reliability, interrater reliability, whether one rater could have assessed alone, and whether the *instrument i*s reliable and valid. The instrument uses a five-point Lickert continuous (interval) scale, with 1438 possible response items per rater. Tests of rater reliability include rater consensus and consistency; intraclass correlation, employing tests applicable to continuous variables (Pearson product-moment correlation); and interrater reliability (five tests using paired samples and one test of all raters). Raters were found to be reliable. Whether any one of the raters could have been the sole responder was considered by assessing how the raters' responses correlated. One rater's responses correlated most highly with those of the other raters; this rater responded to the most statements, and could have rated the instrument alone. Instrument reliability (internal consistency) was assessed through intraclass consistency using Cronbach's alpha and intraclass correlation coefficient. The instrument was found to be reliable. Instrument validity was determined by construct validity (intraclass correlation) and content validity (having experts complete the instrument). The instrument found to be valid but limited validity testing could be done because the instrument is new. Nonetheless, the instrument can now be used to assess the five case studies. It could possibly be used to assess the other 154 innovations of this government and potentially other government innovations and their organizations.

Introduction

The rate of failure of innovations and their organizations in the public sector is not known; neither is why public sector innovations (PSI) succeed. Are there specific antecedents or

[1] Some of the material presented in this paper has been previously published in Glor (2014a, b; 2015); Glor and Ewart, 2016.

[2] Many thanks to Hugh McCague of the Statistical Consulting Centre, York University, Toronto, for his consultation on this paper. The author is nonetheless responsible for the contents.

factors that facilitate creation, successful implementation and survival or mortality of innovations? Some antecedents of PSI have been identified (e.g. Bernier, Hafsi and Deschamps, 2015; Walker, 2003, 2008). Antecedents have been summarized for dissemination of policy/program innovations by Berry and Berry (2013) and for organizations by Glor (2013) but this does not present a complete picture of the antecedents that precede or are associated with creation or adoption of innovations or their organizations. We know very little about the factors implicated in their success or failure. The organizations are probably coupled with the fate of their innovations, but how tightly? Paper I (Glor, 2017) in this series of papers reviewed the previous research on determinants of creation and disappearance of public sector innovation but found they are not fully understood. That paper provided a copy of the instrument developed to study these issues. An attempt is made in the current, second paper to verify that instrument.

Considerable attention has been given to factors influencing the adoption/dissemination of *policy/program* innovations among the 50 American states (summarized in Berry and Berry, 2013). Berry and Berry identified the factors involved as political, economic and social factors. Studying factors influencing the adoption of *administrative innovations* in Canada, including the GoS, 1995 to 2011, Bernier, Hafsi and Deschamps (2015) demonstrated that ***external (environmental/contextual) factors*** were important to the Canadian federal and provincial administrative innovations submitted for consideration to the innovation award of the Institute of Public Administration of Canada. Using a database of 1563 administrative innovation nominations (page 840), Bernier et al. found the following environmental factors had a significant positive correlation with innovations submitted during the new public management era: high unemployment rate, large government size and majority government. They rejected ideology, high unemployment rate as an indicator of strength of the economy, government slack resources as measured by budgetary surplus, and public investments in research and development (at least in the short term). Walker (2003, 2008), on the other hand, concluded that ***internal antecedents were more important than external antecedents*** in local governments. In this instrument, both external (some the same ones described) and internal factors are studied.

While internal and external antecedents of the introduction of innovations have been explored somewhat, there has been little examination of the factors influencing the full implementation, survival or mortality of innovations. This is especially true for the first few introductions of public sector innovations, which is probably the most risky phase for innovations. Their antecedents are little studied. This article takes a next step in filling this gap by attempting to verify the new empirical instrument (Glor, 2017) identifying factors influencing the creation, implementation and fate of public sector innovations and their organizations.

The testing is based on the results of three raters examining five policy/program innovations and their organizations, a full sub-population of innovations in the GoS. Glor's (2014a) framework was employed to develop the instrument. The instrument consists of four questionnaires with 1438 items (statements) prepared to assess factors potentially influencing the creation, implementation, adoption and/or fate of public sector innovations/organizations.

The three expert raters responded to as many statements as they could. All five income security innovations of an innovative GoS (Glor, 1997, 2002) were examined. The factors were considered twice, once at the time of creation and again 10-15 years later, when four of the

innovations were abolished. The potential external factors examined were ideology; politics; external support (interest group and public support); state of the economy and government finances. The potential internal factors studied were fiscal situation, resources accessed, internal support (administrative support, employee support and full implementation); orders of change; whether an efficacious program model was used; and some effects of the innovations. This paper attempts to verify the raters and the instrument. It considers the reliability and interrater reliability of the raters, whether one rater might have been sufficient, and the reliability and validity of the instrument. The Guidelines for Reporting Reliability and Agreement Studies (Kottner et al., 2011) are followed.

The only government reported in the literature for which all of the innovations have been identified is the GoS, 1971-82 (Glor, 1997, 2002). The instrument is tested on a subpopulation of five of its 159 innovations, its income security innovations. The paper identifies a research framework; discusses the case studies; describes the instrument; outlines the methods, measures and null hypotheses tested; analyzes rater reliability and interrater reliability; assesses whether one rater could have assessed the case studies; evaluates the instrument for reliability and validity; describes the results of the tests used and their significance; discusses the results; and identifies possible future research using the instrument.

Definitions. Innovations are defined as the first, second or third introduction of a new policy, program or administrative improvement in Canada or the USA (the GoS's community) (Glor, 1997: 4, based on Walker, 1969; Rogers, 1995). Damanpour and Schneider (2009: 497) used a somewhat similar definition. An *organization* is defined as "a group of people working together for common causes that are registered or captured as an organization in a reliable organizational population database" (Glor, 2013: 3), in this case, the GoS's budget *Estimates*. A government *community* is the group to which the government compares itself and/or with which it works; in the GoS case, the Canadian provincial governments, the federal Government of Canada, and American state and federal governments.

Research Framework. Glor's research framework (2014a, b) is used to frame this paper. The framework (2014a) recommends studying the fate of innovations and their organizations four different ways: interpretive, humanist, functionalist, and structuralist. An *interpretive approach* considers case studies, preferably matched with case studies of normal organizations (qualitative comparative analysis) (Strauss and Corbin, 1998). A normal organization introduces a few innovations but not many. A *humanist approach* focuses on employees, e.g. managers (Damanpour and Schneider, 2006, 2009), employees (Torugsa and Arundel, 2016), employees who implemented the innovations, how the innovations and organizations affected them and how they affected the innovations and organizations. A *functionalist* approach, the most used, identifies and explores factors correlating highly with increased innovation and organizational mortality. A *structural* approach focuses on the fate of structures—including innovations and innovating organizations—and their demography, measured by founding and mortality rates (Glor, 2014a). A multi-theory approach permits consideration of case studies and effects on people, functions, structures and populations (including demographics). Most studies have employed only one or two approaches but employing more approaches should create better understanding. The five cases are not a sample, but rather the full sub-population of income security innovations introduced by the GoS. The reasons they were chosen are discussed later.

The Study

The three raters were drawn from a ***rater population*** that consisted of those who are well informed about these innovations and their organizations. The population consisted of appointed public servants knowledgeable about one or more of the programs. The Premier and ministers of Social Services and Workers Compensation Board (WCB) 1971-82 are deceased. Public servants, Premier and ministers of the Devine government of 1982-91 were not available to this researcher.

The three ***raters*** are experts on one or more innovations and organizations or other aspects of the innovations and organizations. Rater 1 (R1) worked as an officer and supervisor in Treasury Board (the budget, revenue and management department), Executive Council (the Premier's department) and Health, was the budget analyst for Social Services in 1977-78, and worked for the Devine government for two years. Rater 2 (R2) was a former associate deputy minister of SS and only worked for the Blakeney government (the Devine government cancelled all Order-in-Council appointments shortly after coming to power). Rater 3 (R3) was an officer and then manager who worked for the organization delivering one of the innovations during the Blakeney and Devine governments. As a result, although all three were public servants, they had different perspectives on the innovations and the organizations, thus bringing more information to bear. A rater could not be found who knew the WCB innovation well, but R1 had some knowledge from presenting the innovation to Cabinet. The three raters' capacities to respond varied from some to nearly all of the statements. The instrument was prepared by R1. Written guidelines, also prepared by R1, were provided to R2 and R3. In the field of questionnaire development, without guidelines, ratings are thought to be more likely to drift towards what is expected by the rater. Retests and retraining can address this issue, but were not done in this study as the instrument was only completed once by each rater. The three raters assessed the statements independently. While R1 knew who the other raters were, R2 and R3 did not know.

The ***instrument*** consists of questionnaires exploring factors thought to have potentially affected the creation, implementation and fate of the innovations and their organizations. Two questionnaires address innovations and two others address organizations. The instrument is published in Glor (2017). The verification tests use the entire data set, not sub-sets of the data or questionnaires individually.

Case Studies. The innovations studied were highly innovative categorical income subsidy programs (in the Canadian and American context), introduced by the GoS 1971-82 (Glor, 1997). They included: (1) Cost-shared generously-subsidized, income-tested day care subsidies (covering low and middle-low income families), (2) Family Income Plan (FIP), a subsidy for low income working families with children; (3) Seniors Income Support Program (SIP), a subsidy for low income seniors; (4) Employment Support Program (ESP), providing funding for jobs and work support to people on welfare; and (5) the Workers Compensation Board (WCB) introduction of a regular income component. These innovations are considered a sub-population because they are all of the income security innovations introduced by the GoS. The innovations are thus not a representative sample of all of the innovations but are representative of the problems researchers would face doing a larger study to illuminate the

factors influencing and the fate of the population of innovations. The SS organizations changed over time, due to reorganizations of departmental functions. Their history is outlined in Glor and Ewart (2016). Whether WCB created an organization to manage its innovation could not be determined (SS paid for much of the program).

After more than 40 years, there were ***challenges*** assessing the factors, especially: (1) finding and limited access to key documents; and (2) identifying and determining the importance of changes to innovations and organizations, and (3) identifying their fate. Often programs kept very similar names yet changed fundamentally (e.g. objective, basis for determining eligibility). This study uses the *criteria for survival and mortality* outlined in Glor (2013): Creation of an innovation is appearance in the budgetary Estimates or other official document, usually an annual report. Mortality is disappearance from the GoS' budgetary estimates or other official document or a name change in the estimates. The four Social Services innovations and their organizations were abolished during the 1980s; the WCB innovation survived, and continues to this day. The SS organizations were reorganized during the 1970s and again after 1982, then disappeared during the 1980s.

Method and Measures

Using accessible documents,[3] personal knowledge,[4] and creating descriptive statistics, earlier research (Glor, 1997, 2001, 2015; Glor and Ewart, 2016) identified a range of ***possible factors*** influencing the creation, implementation and abolition of the five innovations and their organizations. The next steps were to develop hypotheses (Glor, 2015) and to develop an instrument to score these factors for the two governments studied (Glor, 2017). In this paper the new instrument assessing antecedents and other factors potentially influencing them and their fates is verified: Rater reliability and instrument reliability and validity are tested.

Three raters completed the instrument, indicating degrees of agreement or disagreement with statements about the potential factors, distributed on a five-point Lickert scale. The scoring was strongly disagree=1, agree=2, neither agree nor disagree=3, agree=4, and strongly agree=5. A higher score indicated that the item being measured was more strongly at work, a score of three of middle strength, and lower scores indicated the item was of low influence. The five choices of variate offered to raters for each statement are discrete choices (1, 2, 3, 4, and 5); despite the discrete scoring, the data derived is treated as continuous, because it is assumed that the three respondents rounded off their responses to the nearest whole. For only one variate, one rater did not. Besides identifying strong and weak influences, a higher score is thought to identify an innovation/organization more likely to survive; a lower score to identify an innovation/organization more likely to disappear. Because the measurements were continuous (quantitative, interval), interrater reliability was able to be assessed using paired *t*-tests.

The study tests the reliability and interrater reliability of the raters, the reliability of each

[3]Only recent documents are available online. Earlier documents are rarely available outside Regina. My thanks to the Sask. Archives for copying a substantial number of documents.

[4]Having worked as Social Services Budget Analyst in the Department of Finance; done a special project on the WCB while there; and having worked on the WCB conversion while in Executive Council.

rater, with a special interest in R1, and the reliability and validity of the instrument on the sub-population of the income security innovations and their organizations in two time frames—when they were created by the Blakeney government and when they were abolished by the next, Devine government (4 case studies) or continued under future governments (1 case study). The tests conducted and their results are summarized in Appendix I.

Rater Reliability

Rater reliability, interrater reliability and the reliability of R1 are examined.

Rater reliability is defined as consistent measurement between and among them raters (Salkind, 2011: 102). There are ***four kinds of reliability***, according to Salkind (2011, Chapter 6): test-retest, parallel forms, internal consistency, and interrater reliability. *Test-retest reliability* examines whether a test is reliable over time. Kottner et al. (2011: 96-7) refer to it as intrarater agreement/reliability, in which the same rater, using the same scale, assesses the same subjects/objects at different times. All five innovations were examined, once, within four months of each other, so this test of reliability is not possible in the current research. *Parallel forms reliability* examines different forms of a test. The instrument tested (Glor, 2017) is the first (and only) version of it, so this kind of reliability was not examined either. Two additional kinds of reliability were tested, internal consistency reliability and interrater reliability.

Internal consistency reliability tests rater consistency by calculating a number of correlations among the raters, including tests applicable to continuous variables[5] (Pearson product-moment correlation) and ordinal variables (Spearman rank-order correlation), and rater consistency (rater reliability). The analysis also assesses whether R1 could have been the sole responder to statements in the instrument.

Interrater agreement is the degree of agreement between two raters. There are ***three ways*** to do it: comparison of official ratings, internal consistency correlations and best ratings. (1) Reliable raters agree with an "official" rating. Because this is a new test, there is no official rating available for comparison, so this test is not done. (2) Reliable raters agree with each other about the exact ratings to be awarded. This test is done as part of the internal consistency. (3) Reliable raters agree among themselves which is the best rating. This is also done. Correlations and interrater reliability tests are used in this research to establish: whether the three raters agree overall on the ratings that should be given to the statements (whether they agree with each other as experts); where they stand on the score continuum of the statements provided and whether there are any outlier raters; whether one rater is the best rater and if so, whether that rater's ratings could have been used to determine the results, alone.

Interrater Reliability. Kottner et al. (2011: 97) suggested that "reliability and agreement are not fixed properties of measurement instruments but rather, are the product of interactions between the instruments, the subjects/objects, and the context of the assessment". They are affected by variability in the measurement setting and the statistical approach, which must be assessed. Such variation is tested here, e.g. variability of the mean, paired *t*-test, intraclass

[5] See https://statistics.laerd.com/statistical-guides/types-of-variable.php

correlation (intersubject variability). Interrater reliability is important because it (1) demonstrates whether independent raters can agree on the meaning and assessment of the statements and can reliably rate the statements provided; (2) identify whether one rater is a more reliable rater and can be relied upon to do all of the ratings; (3) addresses directly the subjective elements of making judgments; and (4) has important implications for the validity of the study results (see later) (Stemler, 2004: 1). Interrater reliability is tested here using a paired *t*-test.

Stemler (2004) distinguished ***three kinds of interrater reliability***: measurement estimates, using all data; consensus estimates; and consistency estimates. *Measurement Estimates* describe tests that develop a *summary score for each rater*. They are based on the assumption that all of the data available from all of the raters should be used in creating a summary score for each rater and are most useful when the levels of the rating scale are meant "to represent different levels of an underlying unidimensional construct" (Stemler, 2004: 5). This is not the case with the instrument being examined in this paper, because of the large differences in number of statements to which the raters responded. This test is therefore not done. Consensus and consistency measures compare items individually and only use data from items that both raters being compared have rated.

Consensus estimates. Stemler (2004) called identical scoring of items "consensus". Others use the expression "*exact/perfect agreement*". In this paper consensus is measured two ways. First, it is tested by frequencies, the number of times the raters agree exactly with each other about the scoring of a statement. Second, the paired *t*-test used in this study tests whether there is zero disagreement between raters, and if not, how close to zero the responses are. It is a test of *variability*, based on measuring the standard deviation and estimating the standard error. The *t*-test compares individual ratings on individual questions and calculates a mean difference for each pair, thus developing an overall score of consensus.

Consistency estimates. Stemler used the term "consistency," others use the term "reliability of raters," for tests of whether the raters score items *similarly*. Consistency is measured here by comparing the frequency with which raters answer questions a similar way (defined as a combination of the same way, one lower Lickert scale unit and one higher unit).

Interrater Reliability. The ***paired t-test*** assesses whether the difference in the ratings of paired raters is zero—whether overall they agreed completely on the scorings. Raters are expected, however, to disagree somewhat. Only in relation to unambiguous measurement and measures and only if they hold exactly the same opinions would they agree fully. Measurements of ambiguous factors are generally improved by securing ratings from multiple, trained raters, but this will not always possible in the current research program, hence it is important to determine how differently multiple raters would likely rate the instrument. Disagreement (sources of error) can be due to such factors as variability in interpretation of measurement instruments, measurement procedures, interpretation of measurement results and differences in knowledge of and experience with the innovations and their organizations. Clearly, stated guidelines for rendering ratings are required. Reliable raters demonstrate their independence by disagreeing slightly. This can be evaluated by the Rasch model, in which the probability of a specified response (e.g. right/wrong answer) is modeled as a function of person and item parameters. Very few of the statements in the current study had right/wrong answers, so this test

was not done. Interrater reliability measures how much two raters agree on their ratings of the statements. Raters are compared two at a time. A paired *t*-test was employed to assure that multiple (three) independent raters were assigning the same scores to the same variables, that is, they were achieving *consensus* on the meaning of the statements and thus how they should be scored overall. The more similar the ratings are, the higher the interrater agreement (higher the reliability).

Typically, interrater reliability provides some assurance that multiple raters can be used, but this is not the use to which it is put in the current study. Rather, this study of five cases tests whether the raters are reliable (consistent in their answers) and whether one rater is a more reliable, compared to other independent and well-informed raters. This is used to determine whether one rater could have been relied upon alone, to assess the income security innovations and organizations and to provide some support for this rater alone assessing some or all the remaining 154 innovations and their organizations.

Subjectivity. As well as objective elements, judging statements has subjective elements, as the ratings given depend upon the raters' interpretations of the statements, their knowledge, and experience. Three strategies were used to understand and reduce subjectivity: (1) a set of written instructions was prepared and used by all three raters; (2) three raters with different experience of and knowledge about the case studies were chosen, thus creating balance in the assessments and a broader scope of knowledge; and (3) an interrater reliability test was conducted.

Whether R1 could be the sole rater is important to this study because it is expected that other raters will not be available if the research considers additional innovations, introduced by the same government. It was central to determine how reliable a rater R1 is, compared to other well-informed raters. R1's reliability was assessed by having two additional raters complete the questionnaires and testing interrater reliability through consensus and consistency.

Verifying the Instrument

The instrument was tested on five innovation case studies and their organizations. It was broken into four questionnaires (Appendix I, II, III, IV in Glor, 2017) to make it more manageable: two for innovations and two for organizations. Appendix I and II assess global (environmental) factors and Appendix III and IV test the more specific factors requiring individual innovation/organization assessments. This was confusing for one of the raters, as identical statements were used to measure two different items. The raters completed the questionnaires for two time periods, both of which appeared in the questionnaires: the period when the innovations and organizations were created, during the Blakeney government and for the period when four were abolished, the Devine government. The WCB innovation continues to exist. Event histories and comparisons of the survival periods for innovations and organizations were published in Glor and Ewart (2016: Table 2, 4), after the raters had completed the questionnaires. In the questionnaires, the items were grouped into possible factors that were given names (e.g. ideology, effects). Time 1 and Time 2 were assessed in the questionnaire in proximity to each other, either in the same item or in items that followed one another. In some cases, the same or similar statements were used to examine both innovations and organizations.

Instrument reliability and validity were tested; they are closely associated. Reliability of an instrument is defined as consistent measurement within it; validity of the instrument is defined as an instrument measuring what it is meant to measure. An instrument cannot be valid unless it is reliable but it can be reliable without being valid (Salkind, 2011: 102-3).

Reliability of the Instrument

A test of ***internal consistency reliability*** was done to determine whether the items on the test examined only one dimension. This test also revealed whether the statements are each examining the same dimension and only one dimension. Consistency (reliability) of the items was tested using ***Cronbach's alpha***, a test that only requires one test administration. Cronbach's alpha measures how consistently individual item scores vary with the total score for the test. The more consistently they vary, the higher the value, which provides more confidence that the instrument (and any scales) are internally consistent and measure one thing (Salkind, 2011: 111). A Cronbach's alpha of 0.7 or more is typically considered acceptable in social science research (http://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/). A high alpha does not imply the measure is unidimensional, however. Dimensionality needs to be tested in other ways e.g. exploratory factor analyses (http://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/). This will be considered when the content of the instrument is analyzed, in future papers.

A test of paired samples–***intraclass consistency reliability***—determined whether measures being used in the instrument were reliable measures. The more similar the ratings, the higher the instrument's intraclass reliability (Salkind, 2011: 115). ***Intraclass correlation*** is a way to compare ratings among three raters at a time and a good way to perform reliability testing with continuous data. It assesses the proportion of variance of an observation due to between-subject variability in the true scores, with a range in the scoring between 0.0 and 1.0. Intraclass correlation is an improvement over Pearson's *r* and Spearman's *p* because it takes into account differences in ratings for *individual segments* as well as the correlation between raters. When the scoring is continuous, as argued here, agreement can also be assessed by considering the agreement of raters two at a time or the difference between each pair of raters' observations. The mean of these differences is *bias* and the *limits of agreement* provide insight into how much random variation may be influencing the ratings. When the raters tend to agree, the differences between the observations will be near zero. When one rater is often higher or lower than the other(s) by a consistent amount, the bias will be different than zero. Confidence levels (95%) are calculated for the bias and each of the limits of agreement.

Validity of the instrument

Validity of an instrument is consistent measurement every time it is used (Salkind, 2011: 102). According to Salkind (2011: 117-121), the most important types of validity are criterion, content and construct validity. *Criterion validity* establishes whether test scores are systematically related to other criteria that measure the same factors and whose validity has already been established. *Content validity (face validity)* establishes whether a sample truly reflects a universe of items on a topic. It is assessed by consulting experts. If the items seem to tap aspects that experts in this field regard as important, the questionnaire is considered to have

content validity. *Construct validity* correlates the total of the scores with a theorized outcome that reflects the construct. This research considered content validity and emphasized construct validity. Each of these is considered here.

No criteria were available to do *criterion availability*. *Content validity* of the instrument was assessed by consulting three experts. The *construct validity of the instrument* was tested by intraclass correlation, which is done within the framework of analysis of variance (ANOVA), to compare the means of more than two groups (Salkind, 2011: 221-25). Statistical significance is determined by a ratio of two variances.[6]

The instrument was tested on the sub-population (all of) the income security innovations created by the Blakeney government, 1971-82. They were created first or second in their community, Canada or the USA. All analyses were carried out using IBM SPSS Statistics 24.

Null Hypotheses. The following null hypotheses were tested:

1. The raters are not reliable: The true mean difference among the ratings by the three raters is not zero (has high variance).
2. The raters disagree with the content of the questionnaires.
3. The instrument has zero intraclass correlation
4. The instrument is not reliable.
5. The instrument is not valid.

Results and Discussion[7]

Rater reliability and instrument reliability and validity were tested.

Rater Reliability

The three raters each rated as many of the 1438 statements as they could. This was both a strength—knowledge was distributed, and a weakness—many statements were not assessed by all raters. It meant fewer statements assessed but a wider breadth of statements assessed. The innovations and their organizations were created between 38 and 44 years ago, and other raters were not available, due to deaths and loss of contact. The raw data is outlined in Table 1.

Because this could likewise be a problem in assessing other innovations and organizations, this study also assessed whether R1 was adequately informed to do all the assessments alone, compared to the two other expert raters. R2 was well informed about Time 1 but not about Time 2 and answered only a few of the questions about Time 2. R3 was well informed about one of the innovations, the Employment Support Program (ESP) and its organization, at both Time 1 and Time 2 but was not as well informed about the other innovations and organizations and did not

[6] It is important that the differences are normally distributed. The T-test produces valid results if off-normal.
[7] The definitions used in this section are paraphrased from Salkind, 2011. Others' definitions, if used, are noted.

answer most of the questions about them. There is therefore a risk that ratings will not be the same because of raters' differences in knowledge, experience and opinion—raters do not fully agree. The fact that R1 developed the questionnaires may also be a factor, choosing issues thought to be important, but also potentially, without realizing it, ones that s/he understood, recognized and could assess (although s/he also did not assess some of the statements). This too makes it important to compare R1 to the other raters.

Table 1: Raw Rater Data

R1

	Ratings	Frequency[8]	Percent	Valid Percent	Cumulative Percent
Valid	1	224	15.6	20.0	20.0
	2	129	9.0	11.5	31.5
	3	52	3.6	4.6	36.1
	4	197	13.7	17.6	53.7
	5	519	36.1	46.3	100.0
	Total	1121	78.0	100.0	
Missing	NA + Blanks	317	22.0		
Total		1438	100.0		

R2

	Ratings	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	21	1.5	6.7	6.7
	2	35	2.4	11.1	17.8
	3	32	2.2	10.2	28.0
	4	134	9.3	42.7	70.7
	5	92	6.4	29.3	100.0
	Total	314	21.8	100.0	
Missing	NA + Blanks	1124	78.2		
Total		1438	100.0		

R3

	Ratings	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	159	11.1	22.5	22.5
	2	123	8.6	17.4	39.9
	3	39	2.7	5.5	45.4
	4	126	8.8	17.8	63.2
	5	260	18.1	36.8	100.0
	Total	707	49.2	100.0	
Missing	NA + Blanks	731	50.8		
Total		1438	100.0		

NA= No Answer

Rater consensus correlations between and among the responses of the raters were therefore calculated. Pairs of raters were compared for the ratings they both did, to assure they were reliable raters for the statements to which they both responded. *Rater consistency correlations* were also assessed, by frequencies, proportion of agreement, and Pearson's product-moment correlation, verified by Spearman rank correlation coefficient. *Rater reliability* was also assessed using Cronbach's alpha and interrater reliability using the paired *t*-test.

[8] Basic/raw agreement, not corrected for chance, which Cronbach's does.

Rater Consensus: Frequencies, percentages, means. Finding response consensus among raters provides evidence that the raters understood the statements, understood them the same way and agreed on how they should be assessed. Using all of the statements answered by each rater, R1 responded to 1121 statements, 78.0 per cent of them; R2 to 314, 21.8 per cent; R3 to 707, 49.2 per cent. The histograms of numbers of each of the five possible responses chosen by each rater (not shown) was bowl-shaped (convex) for R1, with a mean score of 3.59; shifted to the right in a concave upward shape for R2, with a mean of 3.77; and convex for R3, with a mean of 3.29. R1's mean was between the other two. The differences in the numbers and choices of statements to which raters responded is a factor in the differences of means, but so are the differences in the experiences and knowledge of the raters, reflecting such factors as length of time working for the government, which of the two governments, and what information was available to the rater. It is a weakness in the data that all raters could not score all of the statements, but the differences are enlightening.

One-way consensus (complete agreement) was tested by frequencies. The number of times each rater agreed completely with each of the others, two by two, was calculated (Table 2). R1 and R2 responded to 309 of the same statements and responded to them identically 46.6 per cent of the time. R1 and R3 responded to 701 of the same statements, identically 73.5 per cent of the time. R2 and R3 responded to 161 of the same statements, identically 46.4 per cent of the time. The level of agreement between raters 1 and 2 and 2 and 3 concerning how the statements should be scored is lower than agreement between R1 and R3. R2 scored the statements somewhat differently than R1 and R3.

The mean score of each rater, using all of the statements the raters answered, was calculated. R1's mean was again in the middle, at 3.59, R2's 3.77 and R3's 3.29. R2 rated the statements answered slightly higher than did the other raters. This difference was largely due to R1 scoring 13.7 per cent of the statements as "4", while R2 scored 9.3 per cent and R3 8.8 per cent "4". R1 scored 36.1 per cent "5" while R2 scored 6.4 and R3 18 per cent "5". Because of the differences in the numbers of statements assessed, these figures are not as meaningful as might be wished.

The means of the differences among the pairs were calculated. The mean difference of pair R1 minus R2 was 0.19, with a standard deviation (SD) of 1.189, calculated on 309 statements (N). The mean difference of R1-R3 was 0.32, with a SD of 1.101 and an N of 70. The mean difference of R2-R3 was 0.22, with a SD of 1.359 and an N of 161. These differences of means are small, less than 0.5 of a Lickert scale for each pair (Table 2). Later the paired samples *t*-test will be done regarding these mean differences.

Rater Consistency takes into account similar responses as well as identical ones. Consistency was considered two ways: (a) How similar the responses were and the percentage of similar responses; and (b) Each rater's response pattern. Concerning ***similarity of responses***, the responses were quite similar. The percentage of identical responses (score of 0) were combined with differences of 1 score (-1 to +1): consistent scores were 82.6 per cent of R1 minus R2's responses, 84.7 per cent of R1 minus R3's responses, and 75.8 per cent of R2 minus R3's responses (Table 2). This indicates the raters agreed very substantially about what the responses

should be. R1's responses were closely similar to those of R2 and R3. The fit between the responses of R2 and R3 were good but not as good as those of R1 with R2 and R3. Concerning the ***patterns of responses,*** each pair of raters had the same pattern, with consistent agreement on the scoring for most questions, and disagreement on many fewer. R1 minus R2's scoring was only 17.4 per cent different, and R1 minus R3's scoring was only 15.3 percentage points different. R2 minus R3 was a little more different, with 24.2 per cent of answers different. Consistency was high and the patterns of responses similar among the three pairs.

Table 2: Consensus and Consistency

Score Diff-erence	**R1 minus R2 Freq-uency of Difference**	**% of Total**	**Valid %***	**Cu-mu-lative %**	**R1 – R3**	**% of Total**	**Valid %***	**Cu-mu-lative %**	**R2 – R3**	**% of Total**	**Valid %***	**Cu-mu-lative %**
-4	1	0.1	0.3	0.3	5	0.3	0.7	0.7	1	0.1	0.6	0.6
-3	2	0.1	0.6	1.0	4	0.3	0.6	1.3	2	0.1	1.2	1.9
-2	20	1.4	6.5	7.4	5	0.3	0.7	2.0	10	0.7	6.2	8.1
-1	41	2.9	13.3	20.7	31	2.2	4.4	6.4	24	1.7	14.9	23.0
0	144	10.0	**46.5**	67.3	515	73.5	**73.5**	79.9	75	5.2	**46.4**	69.6
1	70	4.9	22.7	90.0	48	3.3	6.8	86.7	23	1.6	14.3	83.9
2	17	1.2	5.5	95.5	33	2.3	4.7	91.4	16	1.1	9.9	93.8
3	10	0.7	3.2	98.7	55	3.8	7.8	99.3	6	0.4	3.7	97.5
4	4	0.3	1.3	100.0	5	0.3	0.7	100.0	4	0.3	2.5	100.0
Total	***309***	***21.5***	***100.0***		***701***	***48.7***	***100.0***		***161***	***11.2***	***100.0***	
Missi ng	***1129***	***78.5***			***737***				***1277***			
Total	***1438***	***100.0***			***1438***				***1438***			
Full Agreement (consensus) (%)			46.5				73.5				46.4	
Similar (consistency) (%)			82.5				84.7				75.8	
Different (%)			17.4				15.3				24.2	

* Valid: Taking into account only statements to which raters responded.

Interrater Reliability

Interrater reliability was calculated by studying two raters at a time, five ways, to determine: (1) Frequency counts for proportions of identical, similar and different responses to individual statements; (2) Correlations between responses (Pearson Product-moment correlation, Spearman rank-order correlation [for data that can be ranked]); (3) Independence (chi square); (4) Whether the results occurred by chance (Kappa statistics); and (5) Whether the two sample means are truly different (paired samples *t*-test).

Frequencies and percentages. How identical (consensus) and how similar (consistency) responses were was discussed in the previous section, studied through frequencies and their percentages. Table 2 shows there was a remarkable amount of consensus (difference = 0) among R1 and R3 (73.5%). R1 and R2 (46.5%) and R2 and R3 (46.4%) had a good level of identical scoring. The scorings were very similar.

Correlations between pairs. The least robust measure of interrater reliability is a *count* of how many times the raters agreed exactly with each other, divided by the number of ratings, to calculate the mean number of agreements. The mean number of full agreements was high and high-moderate among the raters, as indicated in Table 2—0.465, 0.735, and 0.464. Counts assume the data is nominal (not the case in the current research). It also ignores the possibility of agreement by chance. Kappa statistics took account of the amount of agreement that could have been expected due to chance, assuming the data is nominal and not ordered (see Table 5, later). The Lickert scale used in this research produces continuous data, not nominal data, but it is nonetheless possible to use Kappa statistics.

The responses of R1 and R2, R1 and R3 were highly correlated, with *correlation coefficients* of 0.546 and 0.772 (Table 3). R2 and R3 were not as highly correlated but the

Table 3: Paired Samples Pearson Correlation

		R1	R2	R3
R1	Pearson Correlation	1	.546***	.772***
	Sig. (2-tailed)		0.000	0.000
	N	1121	309	701
R2	Pearson Correlation	.546***	1	.419***
	Sig. (2-tailed)	0.000		0.000
	N	309	314	161
R3	Pearson Correlation	.772***	.419***	1
	Sig. (2-tailed)	0.000	0.000	
	N	701	161	707

**. Correlation is significantly different from zero at the 0.001 level (2-tailed).
Note: Duplicate data could have been deleted but this presentation shows total numbers of statements to which each pair responded.

correlation is still relatively large for the social sciences, 0.419. All three correlation coefficients were significantly different from zero at the 0.001 significance level (Table 3). The null hypothesis that the responses among R1, R2, and R3 were uncorrelated can therefore be rejected.

Pearson Product-moment Correlation (Pearson's rho, *r)* (also known as Pearson's correlation/Pearson's correlation coefficient) is a parametric correlation,[9] used to investigate the relationship between two quantitative, continuous variables. It was used to measure the strength of the linear correlation between each of the paired variables. Pearson's *r* is based on the method of covariance and provides information about the magnitude of the association or correlation, as well as the direction of the relationship. Pearson's *r* requires the data being studied to meet the following data assumptions: (1) interval or ratio level, (2) linearly related, (3) bivariate normal distribution. Table 3 outlines the results for all statements to which R1, R2 and R3 responded,

[9] Parametric statistics assume the data has a normal distribution (shape), the sample is large enough to represent the population, the variances of each group are similar, and the statistics are *parameterized* by mean and standard deviation. (Salkind, 2011: 285).

testing the null hypothesis that the *true mean difference* among the ratings by the three raters was not zero. Pearson's *r* assessed consistency of responses and examined the relationship between each of the pairs of raters, assessing overall how close the responses of the three raters were. All statements that a rater scored were included, even though raters did not necessarily respond to the same statements. Correlations are for statements that both raters answered.

In general, the raters *agreed* on the responses to the statements. They also responded *similarly*. As hoped, there was far more agreement than disagreement, tested, for example, with frequencies (Table 2). Because there were differences in the statements to which the raters responded, other tests were also required.

Spearman rank correlation coefficient (*rho, p)* is a nonparametric[10] measure of rank correlation (statistical dependence between the ranking of two variables). While it is usually used for ordinal (discrete) data, *rho* is also applicable to continuous data. *Rho* is defined as the correlation coefficient between the *ranking of two variables* (Salkind, 2011: 294). The results of Spearman rank correlations were similar to those for Pearson's *r*.

Pearson chi square is a nonparametric test that determines whether what is observed in a distribution of frequencies is what would be expected to happen ***by chance*** (Salkind, 2011).

Table 4: Pearson Chi Square

Rater	**Value**	**Degrees of Freedom**	**Asymptotic[11] Significance (2-sided)**
R1 vs R2 Valid Cases	309		
Chi square	189.802	16	0.000
Likelihood Ratio	149.282	16	0.000
Linear-by-Linear Association	91.787	1	0.000
R1 vs R3 Valid Cases	701		
Chi square	1049.764	16	0.000
Likelihood Ratio	896.581	16	0.000
Linear-by-Linear Association	416.658	1	0.000
R2 vs R3 Valid Cases	161		
Chi square	87.776	16	0.000
Likelihood Ratio	84.985	16	0.000
Linear-by-Linear Association	28.056	1	0.000

Note: It would be possible to delete likelihood ratio and linear-by-linear from this table, because it was dealt with under correlation, but they are retained here because this table clearly indicates the 1 to 1 ratios.

[10] *Nonparametric* statistics make no assumptions about the probability distributions of the variables being assessed (Salkind, 2011: 285-95, 434).

[11] A line that continually approaches a given curve but does not meet it at any finite distance.

The null hypothesis is that the consensus and similarities found in this research occurred by chance, not because the raters agreed with each other. Because all of the results were significant at the 0.05 level (Table 4), the null hypothesis can be rejected. The responses are not independent of each other; when one scoring goes up, the other does as well. Correlation dealt with this as well.

Kappa statistics are used to analyze qualitative data, assuming the data is nominal (no order). It is a measure of inter-rater agreement for categorical responses but it also tests whether the agreement found was due to ***chance*** by assuming that when raters do not know an answer, they guess. It can also be used to analyze ordinal data. Kappa is a ratio (proportion) that considers observed agreement with respect to a baseline agreement. ***Three types of Kappa*** were calculated—unweighted Cohen's Kappa and two weighted kappas, Cicchetti-Allison (inverse integer spacing) and Fleiss-Cohen (inverse square spacing). ***Unweighted Cohen's Kappa*** analyzes exact agreement between two raters; it does not measure the degree of disagreement. This is especially relevant when the ratings are ordered, as they are here. To address this problem, two modifications to Cohen's Kappa were developed. ***Weighted Cohen's (Cicchetti-Allison or inverse integer spacing) Kappa*** is especially useful when codes are ordered. ***Weighted Fleiss' (inverse square spacing) Kappa*** analyzes agreement among three or more raters. Weighted kappas allow for some close agreement in addition to exact agreement.

Table 5: Kappa Statistics—Ratios of Agreement

	Matrix 1: R1 vs R2		Matrix 2: R1 vs R3		Matrix 3: R2 vs R3	
N of Valid Cases	309		701		161	
N of Missing Cases T=1438	1129		737		1277	
Type of Kappa:	**Score**	**Interpretation**	**Score**	**Interpretation**	**Score**	**Interpretation**
Unweighted Cohen's	.258799	Fair	.635024	Substantial	.256990	Fair
Approx. Significance of Unweighted*	0.000		0.000		0.000	
Weighted Inverse Integer**	.406730	Fair, nearly moderate	.702662	Substantial	.343157	Fair
Weighted Fleiss-Cohen Inverse Square***	.536806	Moderate	.756659	Substantial	.411721	Moderate

* Takes account of closeness. Inverse integer spacing. Not assuming the null hypothesis.
** Weighted kappas would also be significantly different from zero at least at the alpha equals .05 level.
*** Takes account of closeness. Inverse square: difference of one unit is closer than 2 x two units away, like heat from a flame. Fleiss Cohen is a better measure. It assumes the null hypothesis. Categories of Cohen's Kappa are available at: https://en.wikipedia.org/wiki/Cohen%27s_kappa#Significance_and_magnitude (Landis and Koch, 1977; Fleiss, 1981)

The results for these different types of kappas are outlined in Table 5.[12] All three types of kappa tests can be interpreted similarly on a scale of 0 to 1. The scores were fair, moderate and substantial. Although two of the three ratios calculated for unweighted kappa were quite low, this measure requires consensus. The low ratios are therefore still very good. The other two measures are based on consistency (similarity) and so score higher. Comparing all agreement among all three raters, Fleiss-Cohen (weighted inverse square), the best measure here, found moderate and substantial agreement among the three pairs.

Paired Samples T-test. Because of this finding, it is important to do a *t*-test to determine whether the two sample means are truly different. In this case, each rater's set of assessments is considered a sample. A paired *t*-test can be used to compare two population means (in this case sub-population means) when two samples of observations can be paired with observations in the other sample (Mathematics Learning, 2017). There are four conditions for employing a paired *t*-test to compare the responses of the three raters are independence, quantitative measures, conditions of normality, and equal variance (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4667138/). The raters can be assumed to be *independent*: R1 rated the questions first, then chose two other raters to assess the statements. R1 knew who the other raters were, but responded to the statements before identifying and approaching the other raters. R2 and R3 were not aware of the identity of the other raters. The measures are considered to be *quantitative* because the Lickert scale measures qualitative measures quantitatively, on a continuous scale from 1 to 5. Because there are so many values being summed up and averaged (>100 pairs) in this study, their means tend to be normally distributed (central limit theorem). In addition, bootstrapping acts as a check on *normality*, because it does not rely on normality: if the *t*-test and the bootstrapped *t*-test have similar results, this supports the assumption of normality. Finding a significant *t*-test indicates the evidence is sufficient to warrant rejection of the null hypothesis that there is *variability of attitudes* (Ferguson, 1959: 184).

Table 6a: Paired Samples Statistics

Pair	Raters	Mean	N	Std. Deviation	Std. Error of Mean
Pair 1	R1	3.97	309	1.305	0.074
	R2	3.78	309	1.180	0.067
Pair 2	R1	3.61	701	1.634	0.062
	R3	3.29	701	1.624	0.061
Pair 3	R2	3.67	161	1.218	0.096
	R3	3.45	161	1.299	0.102

Table 6b: Paired Samples Correlations*

		N	Correlation	Sig.
Pair 1	R1 & R2	309	0.546	0.000
Pair 2	R1 & R3	701	0.772	0.000
Pair 3	R2 & R3	161	0.419	0.000

* See Table 3.

First, the means of the scores for all statements pertaining to all innovations and

[12] Measuring the kappas: https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english.

organizations for each rater were compared (Table 6a). Second, the paired samples correlations between the means of each pair, considering only the statements they both assessed, were calculated (Table 6b). The correlation between R1 and R3 was high (0.772). The correlation between R1 and R2 (0.546) and R2 and R3 (0.419) were moderately high. R2 responded a little differently from R1 and R3 but all results were significant at the 0.001 level, indicating that the null hypothesis could be rejected. Third, the paired samples *t*-test was conducted (Table 7a). The standard errors (standard deviations) of the means were 0.1 and lower. All results were significant at the 0.05 level, indicating that the paired means are not significantly different.

Table 7a: Paired Samples *T*-Test

		Paired Differences							
	Difference of Means	Mean Difference	Std. Deviation	Std. Error of Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	R1 - R2	0.191	1.189	0.068	0.058	0.324	2.822	308	0.005
Pair 2	R1 - R3	0.322	1.101	0.042	0.241	0.404	7.750	700	0.000
Pair 3	R2 - R3	0.217	1.359	0.107	0.006	0.429	2.030	160	0.044

Table 7b: Bootstrap for Paired Samples Test of Equality of Means

				Sig. (2-tailed) p-value	95% Confidence Interval	
	Dif	Mean Difference	Std. Error		Lower	Upper
Pair 1	R1 - R2	0.298	0.091	0.002	0.118	0.472
Pair 2	R1 - R3	0.516	0.101	0.000	0.311	0.708
Pair 3	R2 - R3	0.217	0.104	0.042	0.012	0.429

The paired samples t-test compared the results on each item (statement) for two raters at a time. It is based on the mean difference between each set of raters, standard deviation, and standard error of the mean. Significance is based on the 95 per cent confidence interval of the difference, 2-tailed. The data could only be compared for questions answered by both members of the pair, minus one degree of freedom to reduce bias. Stemler (2004: 1) indicated that paired *t*-tests should not be over-interpreted, so if other raters employed these questionnaires at other times, their interrater reliability should also be tested.

There are at least three reasons raters would rate statements differently. One is the reason being assessed here, that they were not agreed on the reality assessed in the statements. Other reasons interfere with getting an accurate assessment of this reason. Raters may perceive the statements similarly but one rater may generally score the statements higher or lower than the others. The differences in the means of the scores of Pair 1 (Table 7a) could be due to this. This possibility is tested by the *t*-test. Another possible reason is that this group of raters is not rating in a fully representative way, that one or more of them is an outlier is his/her understanding of the statements being assessed. This is tested and accounted for by standard error, the difference between any one rating and the mean of the other ratings. The standard error gets smaller as the

N get larger. The standard errors here are low. The minimum difference between raters is 1 and the maximum is 4. The standard deviation is close to 1, so the differences are reasonably small.

The results of the *t*-test for each pair are presented in Table 7a. R1 compared to R2 and R3 had significant *p*-values at the 0.01 level with a 95 per cent confidence level that the means are not different. For R2 compared to R3, significance was less, but still significant at the 0.05 level. This is in part because the number of statements considered for the latter test was lower (161), about half the number for the next lower number of statements considered. Statisticians like to see a minimum of 100 items for a strong analysis, so these numbers are fully adequate.

Comparing the mean differences in the scores assessed by paired raters (Table 7a) determined: (1) Whether there was a significant difference in how the raters responded to the statements. On each test, the probability was less than 5 per cent that the test of the null hypothesis that the groups differed was correct. As a result, the ***raters*** were reliable. All differences were found to be excellent[13] (low). (2) Whether the statements tended to be answered the same way (consistently). If they did, the *statements* were reliable. This is addressed in the next section of the paper.

Bootstrapping is a way to ensure that analytical models are reliable, will produce accurate results and derive robust estimates of standard errors and confidence intervals for estimates, e.g. mean and correlation coefficient (http://www.sussex.ac.uk/its/pdfs/SPSS_Bootstrapping_22.pdf). It does not require the data to be normal. If *confidence levels* are similar between non-bootstrapped and bootstrapped methods, it is reasonable to assume the data is normal. The results were similar, so this assumption is reasonable here. Data was bootstrapped for the paired samples *t*-test (based on mean, bias, standard error, bootstrap for 95 per cent confidence level) (Table 7b). Bootstrapping slightly improved the significance of the paired *t*-test for pairs 1 and 3; it did not affect significance for Pair 2 or Pair 3, which remained at 0.000 (Tables 7a, b). It did not change the results as the 0.05 significance level is the usual cut-off for significance. Bootstrapping determined that assuming the data was normal did not change the results.

The *t*-test comparing R1 and R2, R1 and R3, and R2 and R3 were all significant. Although the *t*-test comparing R2 and R3 was less significant, the difference in the mean between R2 and R3 was actually smaller than those of the other pairs (Table 7a, b). It was the smaller number of items that led to a slightly less significant *t*-test.

Given that the scorings of the raters were highly correlated (Table 3) and that the *t*-tests for all three pairs were significant, the null hypothesis that the scorings of the raters were different can be rejected. The scorings of the raters had a high degree of agreement/were reliable.

Reliability of Rater 1

One of the purposes of this research was to determine whether R1 was a sufficiently good rater that s/he alone could have assessed this sub-population and provide evidence for whether

[13] An excellent difference was defined as a difference of less than (<) 1 point in either direction, a good difference as greater than one (>1) point in either direction, an acceptable difference as >1.5 points, and an unacceptable interrater reliability difference as >2 points. One unit was the smallest unit that was offered in the questionnaires.

R1 could then be the sole responder in future research on these governments. If so, this would greatly reduce the amount of research that would need to be done, by eliminating the need for additional raters and interrater reliability tests. In such a case, new raters would probably need to be found for each type of innovation, potentially creating comparability problems. The reliability of R1 is thus an important question for the viability of future research on factors, although not on fate, which can be determined other ways. The criterion for determining that R1 could rate alone required that R1 be as knowledgeable as the other raters. The reliability of R1 was assessed four ways: whether R1 was well informed, whether the means of the three raters' assessments were equal, whether the raters' assessments correlated well, and the consistency of R1's ratings.

Well Informed. R1 responded to more statements than the other raters and unlike the other raters, responded to almost all of the statements (1121 of 1438 statements, 78%, Table 1). R1 responded to 808 more statements than R2 and 419 more statements than R3. R1 could thus be seen as the best informed rater. This assumes that the raters conscientiously answered the questions and did not guess—a reasonable assumption here.

Means. Compared on statements to which both responded, at an absolute level, R1 had a very slightly higher mean than R2 and a somewhat higher mean than R3. R1 had a tendency to rate slightly higher than both R2 and R3. R2 had a somewhat higher mean than R3 (Table 6a). Because of this difference, it was important to consider the correlations between the means of the pairs. They were highly and significantly correlated at the 0.000 level (Table 6b).

The mean of R1 minus the mean of R2 had the smallest difference of mean, but bootstrapped had the second smallest. R1 minus R3 had the largest difference in both cases (Table 7a, b), differences that were significant at the 0.05 level. When bootstrapped (Table 7b), R1 had a slightly higher mean than R2. The difference in the mean responses of R1 and R2 and R2 and R3 were significantly different from zero difference. R1 therefore had a tendency to rate slightly higher than R2 and R3. The difference in the mean responses of R2 and R3 was 0.217, also significant. Nonetheless, the raters scored similarly, because, before the testing, a good agreement was defined as 2 points in either direction and an excellent agreement as 1 or less point in either direction, so the fit of the scores among the raters is excellent. This difference was also examined with a *t*-test.

The paired samples *t*-test determines whether two means are different and provides a score for amount of homogeneity in ratings between two raters (judges) by testing the null hypothesis that the mean difference is not zero. While agreement can occur by chance when only a few categories are being tested, this effect is unlikely here—a large number of items were tested. Interrater reliability was used to assess whether the three judges agreed or disagreed, whether their ratings of the statements were similar, and to determine whether R1 rated the statements the same, a similar or a different way compared to R 2 and R 3. The paired samples *t*-test allowed rejection of the null hypothesis that the raters were not highly correlated because the means were statistically equal and each significance test was less than .05.

Correlations. R1's results were highly correlated with those of R3, who also responded to more statements than R2. The analysis of all three pairs is significant at the 0.000 level and the responses of all three pairs are correlated at the 0.000 level (Table 3). These results allowed

rejection of the null hypothesis that the responses of the raters were different. In other words, the differences among the means are not important differences. The correlations between R1 and each of the other raters were the highest of the correlations, so R1's responses were also most similar to those of the other raters for the statements they both assessed. As a result, R1 is considered the most reliable rater.

Consistency. Cronbach's alpha (Table 8, later) confirmed R1 was the most consistent rater, and that his/her ratings were consistent with those of the other raters.

Sole Rater for Future Research. Because R1 is the most reliable of the raters, R1 could have been the sole rater for these five innovations and their organizations. If the research expands to include the other 154 innovations, and other types of innovations, R1 is also likely to be a better rater to study them. It probably will not be possible to find other raters for other innovations, so the finding that R1 is a reliable rater and the most reliable rater provides some assurance that R1 could be a reliable sole rater for the other innovations and their organizations. This may be a necessity rather than a choice in further research, so this finding is reassuring.

There is a weakness in concluding that R1 could be the sole rater for the portion that can be studied of the remaining innovations. This weakness is based on the fact that the innovations in the pilot were all of the income security innovations introduced; consequently, none of the remaining innovations will be income security innovations. So, while R1 could have been the sole rater for the income security innovations and organizations, s/he may not be as good a rater for the other innovations. At the same time, it must be asked whether s/he would be a better rater than raters 2 and 3, and whether s/he will be the best rater available. There is no evidence that R2 and R3 would be better raters for the other 154 innovations and their organizations. R1 had a wider range of experience and knowledge, having worked in two central agencies and a (different) line department, and having done budgets and projects related to non-income security innovations. S/he did not, however, work in all of the areas that produced innovations. S/he nevertheless has published two books on the innovations of this government, and done other research about it as well. The other two raters have not done these things, but they have done other things. Thus, while R2 and R3 were good raters for these innovations, they would potentially be worse raters than R1 for most of the remaining innovations. They are also not available. Whether that makes R1 a sufficiently good rater for the remaining researchable topics is the unanswered question, and would need to be addressed as part of that research. A good measure would be how many statements R1 could assess the 154 innovations. Currently, this is a moot point and would only be addressable as part of a larger research project.

The Instrument: Do the Statements Measure the Variables Reliably and Validly?

The previous section explored the correlations among and between and the reliability of the raters. This section considers the reliability and validity of the statements.

Reliability of Statements

To be able to employ the statements in the instrument as measures of the items tentatively

identified, the statements must be "correct representations of the variables measured" (McHugh, 2012: 276). The items were developed from the literature on programs, innovations, organizations, management, the governments involved, and the knowledge, experience and judgments of the author. With so many sources for the measures, the statements were tested for internal consistency, using Cronbach's alpha and intraclass correlation.

Internal Consistency Reliability. Cronbach's alpha was used to determine whether the statements (items) in the test are consistent with each other and embody one dimension (Salkind, 2011: 110). The test considers the consistency between raters to provide assurance that the raters are in simple (not complex) alignment and that the raters have shown similarity in the scores they assigned the statements. Cronbach's alpha correlates the score for each item with the total score for each rater and compares that to the variability for all individual items. Variability is the amount of dispersion or spread in a group of scores. Cronbach's alpha measures how consistently each rater responds to the statements and compares this to other raters (Table 8).

Table 8: Instrument Internal Consistency Reliability: All Variables, Cronbach's Alpha

Raters	**Items**			**Reliability Statistics**	
	Valid N	***N Excluded***	***Total N***	***Cronbach's alpha***	***N of Items***
R1 vs. R2	309*	1129	1438	0.704	2
%	21.5	78.5	100.0		
R1 vs R3	701	737	1438	0.871	2
%	48.7	51.3	100.00		
R2 vs R3	161	1277	1438	0.589	2
%	11.2	88.8	100.0		
R1, R2, R3	161	1277	1438	0.724	3
%	11.2	88.8	100.0		

Abbreviations: N=number, R=rater

* Determined by number of statements to which R2 responded.

Cronbach's alpha is a correlation: A perfect correlation was not expected. Salkind (2011: 85) suggested that in the social sciences a Cronbach's alpha correlation "approaching 0.7 and 0.8 are just about the highest you'll see." Cronbach's alpha is therefore high for R1 compared to R2 at 0.704 and even higher for R1 compared to R3 at 0.871. In other words, R1's responses were highly consistent with those of R2 and R3. The consistency of the responses of R2 and R3 was not poor, but was not as consistent—the Cronbach's alpha of 0.590 was moderately high. This is in part due to R2 and R3 only both responding to 163 statements (200 items give better results). R2 did not respond to many statements for Time 2 and R3 was most familiar with one innovation and organization and not as familiar with the others. S/he did not score many of the statements regarding the other innovations. When R2 and R3 responded, they were very good respondents.

Consistency among all raters, based on responses to 163 statements, was a Cronbach's alpha of 0.696, very high (Salkind, 2011: 85) (Table 8). The responses were highly consistent overall. This suggests that all three of the respondents understood the questions, responded to

them similarly (all that was hoped for), and held consistent opinions about them. The raters were agreed—the instrument was measuring one dimension. Intraclass correlation confirmed consistency overall.

Intraclass correlation coefficient (ICC) is an inferential statistic[14] that assesses the absolute agreement (McGraw and Wong, 1996: 33), consistency and reproducibility of quantitative measurements made on units organized into groups that share metric and variance (https://en.wikipedia.org/wiki/Intraclass_correlation#Interpretation) (McGraw and Wong, 1996: 30). In this case the measurements are made by three observers measuring statements on income security. Intraclass correlations are not calculated on pairs. These ratings only included inter-observer variability: the rater factor and the objects of measurement (statements) were treated as fixed and people effects were treated as random, resulting in a two-way mixed effects model.[15] "The ICC estimates are based on mean squares obtained by applying … ANOVA models to these data" (Nichols, 1998).

Fisher's original intraclass correlation closely resembled the Pearson correlation coefficient. One key difference between the two statistics is that in the intraclass correlation the data are centered and scaled using a pooled mean and standard deviation, whereas in the Pearson correlation each variable is centered and scaled by its own mean and standard deviation. For the intraclass correlation, scaling is pooled because all measurements are of the same quantity (albeit on units in different groups). For example, in a paired data set where each "pair" is a single measurement made for each of two units (e.g. two raters) rather than two different measurements for a single unit (e.g. measuring results for each rater), the intraclass correlation is a more natural measure of association than Pearson's correlation (https://en.wikipedia.org/wiki/Intraclass_correlation#Relationship_to_Pearson.27s_correlation_coefficient).

Intraclass correlation is the degree of relationship between observations made under randomly chosen levels of rater scores. There was absolute agreement in the ratings 46.7 per cent of the time for single measures with a 95 per cent confidence interval between 33.9 and 52.6 per cent of the time. On average, there were consistent ratings 72.4 per cent of the time, with a 95 per cent confidence level between 60.6 and 76.9 per cent. The corresponding *F* test for both these intraclass correlations was significantly different from zero at the 0.001 level in both cases: the scores given to the statements by the raters were highly correlated and the statements can be considered on average highly reliable.

Cicchetti (1994) gave the following often-quoted guidelines for interpretation of

[14] *Statistical inference* is the process of deducing properties of an underlying distribution by analysis of data. It infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population. Inferential statistics can be contrasted with *descriptive statistics*, which are solely concerned with properties of the observed data, and do not assume that the data came from a larger population. Collected December 12, 2017 at: https://en.wikipedia.org/wiki/Intraclass_correlation (Salkind, 2011: inferential: 9-10, 171-74, 433; descriptive: 8-9,432).

[15] Inter-observer (rater) variability refers to systematic differences among the observers, and describes deviations of a particular observer's score that are not part of a systematic difference. This was examined earlier. Collected December 12, 2017 at: https://en.wikipedia.org/wiki/Intraclass_correlation#Use_in_assessing_conformity_among_observers; Gwet, 2014).

intraclass correlation inter-rater agreement measures in psychology:

- Less than 0.40—poor.
- Between 0.40 and 0.59—Fair.
- Between 0.60 and 0.74—Good.
- Between 0.75 and 1.00—Excellent

According to this guideline, within group single measures of intraclass correlation (Table 9) of 0.467 among the three raters in this study is fair and average measures of intraclass correlation of 0.724 is good. The intraclass correlation was fairly high for average measures and lower for single measures, with little variation between the scores given to each item.[16] The null hypothesis that there was substantial variation in the responses of the raters can be rejected. The raters were close in their assessments.

Are the statements reliable? Do the statements secure consistent responses from different raters, i.e. are the statements reliable? The Cronbach's alpha and intraclass correlation tests found the raters were consistent in their assessments, their assessments correlated well and therefore the statements were reliable (Table 8). A test-retest was not feasible in this study as one of the raters had reached his/her limits for contributing to the study. If other researchers use the instrument, they could look at some additional aspects of statement reliability.

Table 9: Intraclass Correlation Coefficient[17]

	Intraclass Correlation[b] N=3	95% Confidence Interval		*F* Test with True Value 0 [18]			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0.467[a] Fair	0.374	0.557	3.629	160	320	0.000
Average Measures	0.724[c] Good	0.642	0.791	3.629	160	320	0.000

Two-way mixed effects model where rater effects are random and measures (Lickert scale) effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency (how closely related a set of items are as a group) definition. The between-measure variance is excluded from the denominator variance because it has been shown to be absent in previous tests.

c. This estimate is computed assuming the interaction effect (column variance) is absent, because it is not estimable otherwise.

Note: Cronbach's alpha for all three raters (0.696) is the same as intraclass correlation for average measures.

[16] It should be noted that psychologists conduct many controlled experiments, so their measures are likely to be higher than the ones found in this study.

[17] Intraclass correlation is typically considered within the framework of ANOVA, more recently within the framework of one of three models—fixed-effects, random-effects or mixed-effects models. Here it is analyzed as a two-way mixed-effects model, containing experimental factors of both fixed and random-effects types, with appropriately different interpretations and analysis for the two types.

[18] An *F*-test for the null hypothesis that two normal populations have the same variance (are uncorrelated) needs to be used with caution as it can be sensitive to the assumption that the variables have this distribution (Collected December 12, 2017 at: https://en.wikipedia.org/wiki/F-test_of_equality_of_variances; Snedecor and Cochran, 1989). The results here indicate that it is extremely unlikely that the results are uncorrelated.

Validity of Statements

Tests of validity assess whether an instrument does what it says it does, in this case, identifies factors that influence the fate of innovations and their organizations. The three most important types of validity are criterion, construct and content validity (Salkind, 2011: 118-121).

Criterion validity assesses whether a test reflects a set of factors in a current or future setting. In a current setting, it is known as *concurrent criterion validity*. Applying Salkind's approach (2004:119-120) of using external judges and correlating their results to the original findings, concurrent validity was established in the current research by having two other judges besides R1 (who constructed the instrument) complete it and correlating their results with those of R1. Raters' responses were highly correlated (Table 3, Table 6b) and average measures were good (Table 9). For criteria focused on the future, *predictive validity* is tested. Salkind recommends correlating measures of success in the present with the same measures in an earlier time period. In the current research, criteria thought to have been implicated in the successful creation, implementation and achieving of positive effects from innovations/organizations created during the 1971-82 GoS were correlated with the survival of the innovation or its organization during the next, 1982-91 government. In other words, this research asked: Can the factors predict creation of the innovations and their organizations at Time 1? Can they predict the fate of the innovations and organizations at Time 2? Predictive validity will be explored in a different essay.

Construct validity is established by demonstrating the validity of the underlying construct or idea behind an instrument (Salkind, 2011: 120-21). Here it could be demonstrated by: an external standard, such as use of multiple highly knowledgeable insiders as raters (done), and a high correlation with items previously found to correlate with survival and a low correlation with items predicted not to correlate with survival of innovations and their organizations (to be done in a future paper).

There is very limited information available on the fate of innovations and their organizations. Except for a meta-analysis of a few small studies of the mortality of innovations and their organizations (Glor, 2015: 75-90; 2017), factors influencing the fate of innovations and their organizations have been little reported. The literature on antecedents referenced earlier considered what preceded creation of innovation but not what influenced its fate. Antecedents identified in the tool for creation of these five innovations will be compared to the antecedents for disappearance and continuance of innovations. Factors influencing the mortality of specific organizations but not specifically innovative ones have been studied, so this provides a comparison to factors that might be influencing the implementation and fate of the organizations. These have been summarized in Glor (2013) and are reflected in the statements in the instrument. A future essay will explore construct validity in more detail.

To address ***content validity***, three experts (the raters) were consulted. The raters were asked for their comments on the statements. All three raters were experts on the income security innovations studied, albeit in different ways and on different aspects (see details earlier). Glor developed and modified the statements; the other two raters were asked for feedback on the

instrument.

Only R2 provided feedback, on eight statements. The author's comments are in italics.

1. Request for definitions of "quickly" and "easily" in statement 27. *The statement was looking for the rater's understanding of this.*
2. Saying statement 111 was based on fact, not opinion.[19]
3. Showing a lack of recognition that statement 120 was asking for responses in two time frames, and had provided boxes for them. *Perhaps this statement should be divided into two statements.*
4. Requesting a definition of "official report" in statement 124. *The statement was trying to tap the rater's knowledge and understanding of this.*
5. Requesting a definition of "much of the marginal resources" in statement 132. *Same comment as No. 1.*
6. Requesting a definition of how long "sufficient information was available to track the innovation," statement 141. *Same comment as No. 1.*
7. Requesting a definition of how long "the innovation retained its funding" in statement 148. *Part of the purpose of this research was to determine whether information is available on how long the innovations were funded, and at what level, so it was impossible to provide a definition.*
8. Pointing out statements 152 and 154 were identical. *This was deliberate, as it was thought to be potentially twice as important as the other factors.*

R2 did not seem to do research to respond to the questions, relying on memory instead. It appears s/he would have had the same problems finding information as the author.

Given the contradictory responses to statement 158, which reads "The innovations did not act as a disincentive to work," it should perhaps be reworded to read: "The innovations acted as a disincentive to work." This would be a problem, however, as the scoring would need to be reversed in just this one case, which the raters would equally need to note.

Appropriateness of Statements (Items). This issue could only be examined in a limited way. The intent had been that if none of the three raters could (did) respond to a statement, the conclusion was drawn that it was not an appropriate statement to measure the factor. If further research made a response possible,[20] the statement was retained, but if no one knew or could find the information, it was not an appropriate statement (e.g. statements requiring knowledge of the organization that administered the WCB innovation).

This research demonstrated, with a few exceptions, that raters could agree upon the meaning of, how to score, and what the correct responses were to the statements. The instrument was found to be reliable and valid (limited testing of validity). Appendix I summarizes the tests and results verifying the raters and the instrument.

[19] The *Directions*, however, had not indicated the instrument was limited to opinions. The author was hoping some raters would have more knowledge or would be able to do more research than s/he could. This was not the case.
[20] Following the completion of the questionnaires by all three raters, R1 secured access to more data from the Estimates and was able to respond to some additional statements.

Conclusion

The verification process conducted for this new instrument (see Glor, 2017) found:

- The raters could respond to many of the statements but not always to the same ones.
- The scores of the raters were highly correlated.
- Rater and interrater reliability was high, barely with one exception.
- Raters understood the meaning of the instrument similarly.
- R1 could have rated the statements alone.
- The new instrument developed to examine the factors influencing the fate of income security innovations and their organizations of the GoS 1971-82 is reliable and valid.

Rater 1 was found to be the most reliable assessor. This issue is of particular interest to future research, where additional innovations of this government could be studied. Up to 44 years after the innovations were created, it will probably not be possible to find raters for many of the other innovations (the author was fortunate to know raters 2 and 3 and to have received their cooperation). Moreover, it would be better to have one rater rate all of the innovations. Rater 1 being a reliable rater, an adequate rater by her/himself, and his/her responses being adequate to identify the factors, simplifies the process of examining further innovations/ organizations of the GoS and would create better consistency in future research.

The next step will be to consider the content of the ratings. Although there is some literature on antecedents of public sector innovation, it has not been possible to draw many firm conclusions about them. Some researchers, for example, have considered ideology or politics as a factor but they have come to contradictory conclusions (e.g. Berry, Ringquist, Fording and Hanson, 1998; Bernier, Hafsi and Deschamps, 2015). With a reliable and valid instrument with which to consider these issues for innovations and organizations and their fates, hopefully some further things can be learned. It will now be possible to examine the three raters' assessments of the five income security innovations and their organizations with more confidence. This will be done in a subsequent paper, in which the statements about innovations will be examined to see if they are successfully identifying and distinguishing the factors/variables being studied, in Time 1 and Time 2, which factors are most important, and which factors predicted fate.

A full paper has been devoted to verifying the raters and the instrument and finding a sole rater, with two objectives: (1) In hopes that further research will be done on the GoS innovations and their organizations; (2) To make the research as transparent as possible, in hopes that other researchers can benefit from and possibly use the same or a similar instrument, thus creating comparable research and building cumulative knowledge of public sector innovation. With this in mind, the tested instrument was published in Glor (2017).

With the exception of the few statements that presented problems for one rater, the high reliability of the raters provides some assurance that the statements are measuring one dimension and that the statements are doing what they were meant to do. Because the statements are both reliable and valid, they can be used with some confidence to examine other income security programs and organizations. With the exception of the few statements devoted exclusively to

income security, consideration can also be given to using the instrument to study the other 154 Saskatchewan innovations and organizations. In future it should be possible to explore additional innovations and organizations of the Blakeney government: It would be useful to see if the same factors were important for them. It may be possible, for example, to assess whether and by how much fate was influenced by the innovations' ideology and/or politics, as opposed to economic and fiscal challenges. An earlier analysis of all of the innovations of the Blakeney government suggested that about a third were social democratic, a third liberal and a third conservative. Whether the instrument could be used to study innovations and organizations in other governments would require some further thought and assessment by their researchers.

About the Author:

Eleanor D. Glor worked as a public servant in the Canadian public sector at four levels of government and has written about public sector innovation for publication since the 1980s. She ran the Innovation Salon, a meeting on public sector innovation, from 1995-2005 and is the publisher and founding editor of *The Innovation Journal: The Public Sector Innovation Journal.* She is Fellow, McLaughlin College, York University, Toronto. Most recently she published *Building Theory of Organizational Innovation, Change, Fitness and Survival;* What Happens to Innovations and Their Organizations (with Garry Ewart); "Innovation and Organizational Survival Research" with Mario A. Rivera, in James D. Ward (Ed.), *Leadership and Change in Public Sector Organizations: Beyond Reform*; and Studying Factors Affecting Creation and Fate of Innovations and their Organizations I: A New Instrument. She edited a book of articles from *TIJ*, entitled *Leading edge research in public sector innovation: Structure, dynamics, values and outcomes*, which is in press with Peter Lang. Eleanor can be reached at: glor@magma.ca

Acknowledgements of former Saskatchewan public servants: Larry Flynn, regional & acting director, ESP 1974-84; Ron Hikel, former Associate Deputy Minister, social development (income security), Manitoba & Associate Deputy Minister, SS, Sask.; Ian Potter, Planning Bureau, Executive Council & Director Social Services Planning Unit, Department of Social Services, Sask; Merran Proctor, former Director, Social Services Planning Unit (personal email to Eleanor Glor from JustCause, July 28, 2014), Department of SS, Sask. She replaced Ian Potter; Toby Stewart, former head of ESP. The author is also a former Sask public servant.

References:

Bernier, Luc, Taïeb Hafsi & Carl Deschamps. 2015. Environmental Determinants of Public Sector Innovation: A Study of Innovation Awards in Canada. *Public Management Review*, 17(6) (December): 834-56. DOI: 10.1080/14719037.2013.867066

Berry, Frances Stokes & William D. Berry. 2013. Innovation and Diffusion Models in Policy Research. Pp. 307-362, Chapter 9, in Paul Sabatier (Ed.), *Theories of the Policy Process.* Boulder, CO: Westview Press.

Berry, William D., Evan J. Ringquist, Richard C. Fording & Russell L. Hanson. 1998. Measuring Citizen and Government Ideology in the American States, 1960-93. *American*

Journal of Political Science, 42(1): 327-48.

Cicchetti, Domenic V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4): 284–290. *doi*:*10.1037/1040-3590.6.4.284*.

Damanpour, F. & M. Schneider. 2006. Phases of the Adoption of Innovation in Organizations: Effects of Environment, Organization and Top Managers. *British Journal of Management*, 17(4): 215-36.

Damanpour, F. & M. Schneider. 2009. Characteristics of Innovation and Innovation Adoption in Public Organizations: Assessing the Role of Managers. *JPART,* 19(4): 495-522.

Ferguson, George A. 1959. *Statistical Analysis in Psychology and Education, 2nd edition.* Toronto, CA, New York, NY: McGraw-Hill Book Company.

Fleiss, J. L. 1981. *Statistical methods for rates and proportions* (second ed.). New York, NY: John Wiley. ISBN 0-471-26370-2.

Glor, Eleanor D. (Ed.). 1997. *Policy Innovation in the Saskatchewan Public Sector, 1971-82.* Toronto, CA: Captus Press.

Glor, Eleanor D. 2001. Key Factors Influencing Innovation in Government. *The Innovation Journal: The Public Sector Innovation Journal*, 6(2), article 1. Collected March 11, 2017 at: http://www.innovation.cc/volumes-issues/vol6-iss2.htm

Glor, Eleanor D. (Ed.). 2002. *Is Innovation a Question of Will or Circumstance? An Exploration of the Innovation Process through the Lens of the Blakeney Government in Saskatchewan, 1971-82*, 2nd edition. *The Innovation Journal: The Public Sector Innovation Journal.* Collected August 26, 2014 at: http://www.innovation.cc/books.htm

Glor, Eleanor D. 2013. Do innovative organisations survive longer than non-innovative organisations? Initial evidence from an empirical study of normal organizations. 2013. *The Innovation Journal: The Public Sector Innovation Journal*, 18(3), article 1. Collected December 20, 2013 at: http://www.innovation.cc/volumes-issues/vol18-no3.htm

Glor, Eleanor D. 2014a. Studying the Impact of Innovation on Organizations, Organizational Populations and Organizational Communities: A Framework for Research. *The Innovation Journal: The Public Sector Innovation Journal*, 19(3): article 1. Collected March 11, 2017 at: http://www.innovation.cc/volumes-issues/vol19-no3.htm

Glor, Eleanor D. 2014b. Proposal for Comparative Research on the Fate of Innovative Public Sector Organizations, a paper presented to the International Research Society for Public Management, April 9-11, 2014, Ottawa, Canada.

Glor, Eleanor D. 2015. *Building Theory of Organizational Innovation, Change, Fitness and*

Survival. Ottawa, CA: The Innovation Journal: The Public Sector Innovation Journal, 20(2): article 1. Accessed May 23, 2015 at: http://www.innovation.cc/books.htm

Glor, Eleanor D. 2017. Studying Factors Affecting Creation and Fate of Innovations and their Organizations – I: A New Instrument. *The Innovation Journal: The Public Sector Innovation Journal*, 22(2), article 1. Collected September 5, 2017 at: http://www.innovation.cc/volumes-issues/vol22-no2.htm

Glor, Eleanor D. & Garry Ewart. 2016. What Happens to Innovations and their Organizations? *The Innovation Journal: The Public Sector Innovation Journal,* 21(3), article 1. http://www.innovation.cc/volumes-issues/vol21-no3.htm

Gwet, Kilem L. 2014. *Handbook of Inter-Rater Reliability, 4th Edition.* Gaithersburg, Md.: Advanced Analytics, LLC. ISBN 978-0970806284.

Kottner, Jan, Mohamed M. Shoukri, Stig Brorson & David L. Steiner. 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64 (January): 96-106. doi: 10.1016/j.jclinepi.2010.03.002.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics,* 33(1): 159–74. doi:10.2307/2529310; JSTOR 2529310; PMID 843571.

Mathematics Learning Support Centre. 2017. Statistics: 1.1 Paired *t*-tests. Collected January 1, 2017 at: http://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf

McGraw, Kenneth O. & S. P. Wong. 1996. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods*, 1(1): 30-46.

McHugh, Mary L. 2012. Interrater reliability: the Kappa Statistic. *Biochemia Medica (Zagreb)*, 22(3) (October): 276-282.

Nichols, David P. 1998. Choosing an Intraclass Correlation Coefficient. SPSS Keywords, No. 67. Collected April 3, 2017 at: http://web1.sph.emory.edu/observeragreement/spss.pdf

Pontius, Robert & Marco Millones. 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32: 4407–29.

Rogers, Everett M. 1995. *Diffusion of Innovations. 4th edition.* New York, NY: The Free Press.

Salkind, Neil J. 2011. *Statistics for People Who (Think They) Hate Statistics*. 4th edition. Los Angeles, Cal: Sage

Strauss, A. & J. Corbin. 1998. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: Sage

Stemler, Steven E. 2004. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research and Evaluation*, 9(4) (March): 1-11.

Torugsa, Nuttaneeya (Ann) & Anthony Arundel. 2016. Complexity of innovation in the public sector: a workgroup level analysis of related factors and outcomes. *Public Management Review*, 18(3): 392-416.

Walker, Jack L. 1969. The Diffusion of Innovations among the American States. *American Political Science Review*, 63(3): 880-99.

Walker, Richard M. 2003. Evidence on the Management of Public Services Innovation. *Public Money and Management,* 23(2) (April): 93-102.

Walker, Richard M. 2008. An Empirical Evaluation of Innovation Types and Organizational and Environmental Characteristics: Towards a Configuration Framework. *Journal of Public Administration Research and Theory*, 18(4): 591-615.

Appendix I: Summary of Verification Process: Tests Conducted and Results

Assess	Type	Test or Statistic	Results	Interpretation
Rater Reliability (*agreement):*				
Rater Reliability	Comparison of rater responses	Raw data, percentages	Table 1	Raters did not respond to all statements, but a wide breadth of statements assessed.
	Correlations	Paired Samples Pearson Correlation	Table 3	All results significant at .000 level
Interrater Reliability	Interrater Agreement: Number and % of times raters score same and different. Table 2.	Consensus Consistency	R1 minus R2 exact agreement: 10%/21.5% = **46.5**% identical. R1 – R2 similar score[21] = 17.8%/21.5%=**82.8** % similar.	Good exact agreement Close similar agreement
		Consensus Consistency	R1 minus R3 exact agreement: 35.8/48.7 = **73.5**% identical. R1 – R3 similar score: 41.3/48.7 = **84.7**% similar	Close exact agreement Close similar agreement
		Consensus Consistency	R2 minus R3 exact agreement: 5.2/11.2 = **46.4**% identical. R2 – R3 similar score: 8.5/11.2= **75.8**% similar	Good exact agreement Close similar agreement
	Independence of Raters	Pearson Chi square (X^2) test of rater independence Table 4	R1vs R2: $X^2(16)=189.8$, p=.000 R1 vs R3: $X^2(16)= 1049.8$, p=.000 R2 vs R3: $X^2(16)= 87.78$, p=.000	Very strong evidence against independence.
	Measure inter-rater agreement for qualitative (categorical) items: compare number of times raters closely agree and disagree[22]	Fleiss-Cohen (inverse square spacing) weighted Kappa (κ) Table 5	R1 and R2: .536806 R1 and R3: .756659 R2 and R3: .411721	Agreement: R1 and R2: moderate R1 and R3: substantial R2 and R3: moderate.
	Quantitative Between paired samples	Paired samples Pearson Product-moment Correlation Coefficient (*r*) Table 3	R1 v. R2 0.546 Significant .000 R1 v. R3 0.772	Can reject the null hypothesis that the raters are not highly correlated. Can reject null hyp.

[21] Similar score is defined as: same score + one measure less + one measure more on the Lickert scale.

[22] This is thought to be a more robust measure than simple percent agreement calculation, since κ takes into account the possibility of the agreement occurring by chance (Collected December 12, 2017 at: https://en.wikipedia.org/wiki/Cohen%27s_kappa; Pontius and Millones, 2011).

Assess	Type	Test or Statistic	Results	Interpretation
		Supported by Spearman rank correlation coefficient (rho) (*p*)	Significant .000 R2 v. R3 0.419 Significant .000	Can reject null hyp.
	Paired samples	Paired samples *t*-test 2-tailed significance Table 6a, 6b	-R1-R2: Mean=0.191, t=2.822, sig. 0.005, bootstrapped sig 0.002. -R1-R3: Mean 0.322, t=7.750, sig. 0.000, bootstrapped sig. 0.000 -R2-R3: Mean 0.217, t=2.030, sig 0.044, bootstrapped sig 0.042	Can reject the null hypothesis that the raters are not highly correlated. Can reject null hypothesis Can reject null hypothesis
Is one rater more reliable than the other two?	Comparison of Means	Based on paired samples *t*-test, rank raters' means Table 7a, 7b	Bootstrapped sig. 2-tailed: R1-R2: .002 R1-R3: .000 R2-R3: .042	On average R1 scored statements 0.243 higher than the other raters: Statistically significant but a small fraction of one unit of a Lickert scale.
		Rank difference between means of each rater	R1 scored statements 0.243 of a score higher than the mean, R3 0.247 unit higher than the mean, R2 0.027 lower than the mean.	On average, R1 tended to score the statements a little higher than the other raters.
Assessing the Instrument:				
Reliability	Intraclass (Internal) Consistency Reliability: All statements. How closely related a set of items are as a group. Considered to be a measure of scale reliability.	Cronbach's alpha: Expected correlations between raters. Not a statistical test. ≥0.8 very good ≥0.60 good. Table 8	Single measures: 0.467 Average measures (better test): 0.724	Fair Good
	Intraclass consistency or reproducibility of quantitative measurements made on units organized into groups (raters). Table 9	Intraclass correlation: degree one variable can be equated to another variable, adding a constant. Amount of variation in the mean values with positive agreement	3 raters compared to mean values: Single measures: 0.433, fair agreement. 3 raters, average values: 0.696, good agreement.	Can reject the null hypothesis that there was substantial variation in the responses of the raters.
		How close are the 3 measures to the average of the 3 measures?	Single measures: 3.293 Sig. 0.000 Average measures: 3.293. Sig. 0.000 3.5 would be exactly mid-point.	There is a good amount of consistency among the 3 raters, singly and on average. Single measures are more useful.

Assess	Type	Test or Statistic	Results	Interpretation
Validity of statements	Content validity	3 experts' opinions of the instrument.	Five requests for definitions from one rater.	No other source of information available-construct development
	Construct validity -Used external standard: multiple highly knowledge-able insiders as raters	Intraclass correlation coefficient	Within group single measures: 0.467 among the three raters Average measures of intraclass correlation: 0.724 Little variation between the scores given to each item.	Fair Good Null hyp of substantial variation in responses of the raters can be rejected. Raters were close in their assessments.