

Research and Theory

Applying aspects of generalizability theory in preliminary validation of the Multifacet Interprofessional Collaboration Model (PINCOM)

Atle Ødegård, Associate Professor, PhD, Institute of Health and Social Sciences, Molde University College, Postbox 2110, 6402 Molde, Norway

Knut A. Hagtvet, Professor, PhD, Department of Psychology, University of Oslo, Postbox 1094 Blindern, 0317 Oslo, Norway

Stål Bjørkly, Professor, Dr. psychol, Institute of Health and Social Sciences, Molde University College, Postbox 2110, 6402 Molde, Norway; Centre for Research and Education in Forensic Psychiatry, Ullevål University Hospital, Oslo, Norway

Correspondence to: Atle Ødegård, Institute of Health and Social Sciences, Molde University College, Postbox 2110, 6402 Molde, Norway, Tel.: 0047 71214014, Fax: 0047 71214100, E-mail: Atle.Odegard@hiMolde.no

Abstract

Objective: Empirical assessment of the Interprofessional Collaboration (IPC) model within the framework of Generalizability Theory (GT).

Design and participants: A multifacet data collection design served the purpose of examining the relationships between observed indicators representing a defined construct. Professionals working with children and adolescents (n=134), in the context of mental health care, completed a 48 item questionnaire addressing 12 aspects of interprofessional collaboration.

Results: Estimated variance components from two G-studies are presented. The relative impact of different sources of variance was estimated 1) for the full design, and 2) for three groups of informants (schools, primary care and specialist services). Differences between groups were found regarding the relative impact of the level — and context facets with respect to perception of IPC.

Conclusions: The methodology of generalizability theory is well suited for data with a complex facet structure as displayed in the present study. We recommend researchers to give domain specifications thorough attention when measuring IPC.

Keywords

interprofessional collaboration, mental health, generalizability theory

Introduction

There are relatively few instruments developed to investigate interprofessional collaboration (IPC) [1] and the ones that do exist are in an early phase of development and testing [1, 2] or concentrate on interprofessional education (IPE) rather than on IPC in clinical practice [3–6]. The literature is characterized by a conceptualization of the phenomenon [7, 8] and this is illustrated by the following examples of main constructs in use within the field: inter-disciplinary, multi-disciplinary, inter-professional, inter-agency, inter-departmental, multi-institutional, cross-disciplinary [9], integrated care [10] and shared learning [11]. This indicates that

there are good reasons to discuss conceptual problems related to the phenomenon of IPC, especially if one attempts to use measurements. Meads and Ashcroft [12] assert that “in setting out the case for collaboration it will be helpful to begin by exploring something of the range of meaning that collaboration can convey” (p.16). This seems especially relevant as it is likely that often the construct interprofessional collaboration or related constructs are used as “buzzwords” [10], i.e. that they trigger attention and immediate recognition.

The primary objective of this paper is to address the issue of measuring IPC as a complex contextual and multifaceted phenomenon by means of generalizability

theory as a methodological framework to this end. It is not given what IPC means for professionals working at different levels in the social- and health care system. Nunnally [13] claims that as a first step in a measurement procedure the researcher should specify the domain of indicators of the construct. Thus any attempt to operationalize a theoretical construct, such as IPC, on the empirical level may be encumbered with errors. For example, without domain specifications, it is difficult to decide to what extent a measure includes irrelevant information or under-represents the constructs and both may represent serious threats against construct validity [14]. In the case of IPC, giving attention to domain specifications will increase the likelihood of clarifying IPC in a given study, thus also reducing the chances of confusion about what is meant by IPC. This obviously seems to be an important issue as there are many constructs in use intended to pin down the essentials of IPC.

This paper is partly based on the data material used in a previously published paper in *International Journal of Integrated Care* [2]. However, new data are provided as two contexts of collaboration are included — internal and external interprofessional collaboration. To our knowledge the methodology (Generalizability Theory — GT) used in the present article is completely new in the study of Interprofessional Collaboration (IPC) and it is suggested that it may enrich our understanding of the phenomenon of IPC. The main advantage of GT in the study of IPC is that it takes into account the complexity of IPC — especially that perception of IPC to a large extent is contextually dependent. Thus the methodology gives new possibilities in the validation processes and represents an interesting alternative to traditional factor analytic approaches [2].

The most important issue in this article is the definitional issue. Considering domain specifications, is an important step towards reliable and valid item generation, since validation is seriously hampered if it is unclear what constructs the items actually represent. In our view this is a major challenge in the study of IPC. Thus, we believe that the use of GT in the present study may give other researchers who are interested in the measurement of similar constructs (such as integrated care) a new approach to explore their data. In the classical theory, factor analysis, and reliability and validity measures are used to describe how good constructs and latent variables are measured, and this may help to improve questionnaires trying to measure latent constructs. The use of GT to model different sources of measurement error may prove to enhance the reliability of questionnaires that measure complex constructs.

We argue that the complexity of IPC as a phenomenon and construct invites for multi-faceted designs. A multifaceted design may be defined as a design having multiple sources of measurement error. Firstly, because the construct to be studied may involve a much larger complexity than assessable by one-facet designs; “defined by one source of measurement error, that is by a single facet” [15 p.3]. Secondly, based on the diversity of constructs used to describe collaboration, there are good reasons to believe that IPC is a phenomenon best described as a multifaceted phenomenon. IPC has a “build-in” complexity that needs to be explored before reliable and valid scores are obtained. Reliability and validity are basic to any measurement approach. The extent to which a measure actually measures the trait or dimension it is supposed to measure is defined as validity. PINCOM is developed as a preliminary model to increase the likelihood of producing valid scores of IPC. Yet, lack of acceptable levels of reliability render high validity estimates almost worthless. According to Messic [14 p.741] validity is not to be regarded as “a property of the test or the assessment as such, but rather of the meaning of the test scores”. To address test score interpretations of the measurement, it is thereby necessary to provide a framework for these interpretations. Thus, according to Messic [14], validity is to be understood as a judgement of the degree to which empirical evidence and theoretical rationales support the appropriateness of test score interpretations.

Developing a measurement procedure of IPC

Our study reports on the application of a questionnaire, which is intended to measure professionals’ perception of IPC based on their own experience. The questionnaire is at an early stage of development and our intention is to assess the complexity of IPC through a series of estimates of different sources of score variation across multiple contexts and facets pertaining to IPC.

In general, Hinkin [16] asserts, that scale development, for example the development of a questionnaire, should follow three stages: stage 1 — item generation, stage 2 — scale development and stage 3 — reliability assessment. There are several ways to explore the development of measurement procedures and it seems that the most common approach is the use of factor analysis and estimation of reliability [2, 16, 17]. Ødegård [2] has previously emphasized that exploratory factor analysis and reliability testing according to classical test theory may identify the most central subscales in the same questionnaire as used in this study.

However, the development of any measurement procedure should be approached by several theoretical methodologies and statistical analyses, to avoid too early closure. Messick [14], for example, claims that validity is an evolving property and validation a continuing process.

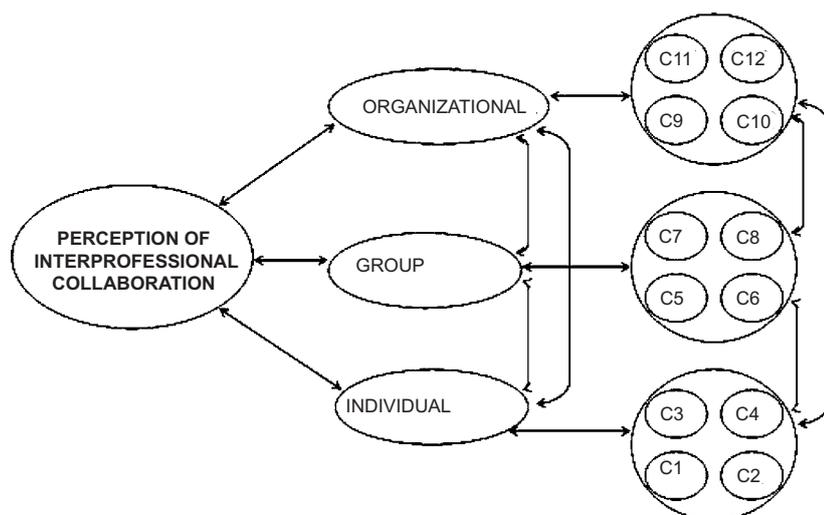
Hagtvet and Zou [18 p.50] claim that “Measurement research and conceptual development should gradually contribute to improving the definition of the construct, which in turn would suggest more precise definition of the corresponding measurement domain”. In the case of IPC this seems very important, as presently there are many concepts and theoretical models trying to grasp the phenomenon of IPC. Possibly the existence of different IPC models [7] makes the development of measurement procedures within the field rather challenging. In two relatively new review studies, San Martin-Rodriguez and co-workers [8] and D’Amour and co-workers [19] claim that the international situation regarding IPC is characterized by a lack of empirical studies; “In addition to the general absence of empirical studies evaluating the impact of various factors on collaboration, some of the studies we have are limited in both their scope and in the methods employed” [8 p.144]. However, researchers striving to measure IPC one way or the other also challenge existing definitions of IPC, because the development of measurement procedures demand precise definitions of the measurement domain. In this study a preliminary attempt to define central features of the IPC construct was

approached by the development of a theoretical model denoted as the Perception of Interprofessional Collaboration Model (PINCOM), previously described by Ødegård [2].

Conceptual framework – the perception of the Interprofessional Collaboration Model (PINCOM)

In the present study, domain specifications were attended to increase the likelihood that the questionnaire reflects central aspects of collaboration within service delivery and case work. However, although deductive in nature, the PINCOM is at an early stage of a priori and this calls for an explorative approach. In this regard the model, and the empirical investigation of the observables to be presented later in this paper, must be considered as an early step in the validation process as validity is a “matter of degree rather than an all-or-none property, and validation is an unending process” [13 p.87].

PINCOM is based on two main sources of information: a) a pilot study [20] and b) relevant literature in the interface between IPC, organizational and social psychology. The tentative theoretical model developed consisted of twelve aspects (sub-constructs) of collaboration within three contextual levels, all of which were assumed to have relevance for IPC (Figure 1).



C1=motivation, C2=role expectancy, C3=personality style, C4=professional power, C5=group leadership, C6=coping, C7=communication, C8=social support, C9=organizational culture, C10=organizational goal, C11=organizational domain and C12=organizational environment

Figure 1. The Perception of Interprofessional Collaboration Model (PINCOM).

Generalizability theory (GT) as a methodological approach in the study of IPC

The IPC construct (*perception of IPC*) presented in the present study, and the sub-constructs included in the PINCOM (C1–C12), suggests that IPC is a multifaceted construct that fits the use of Generalizability Theory (GT). GT is a statistical theory about the dependability of behavioural measurements [15], based on the work of Cronbach and others [21]. “Dependability, then, refers to the accuracy of generalizing from a person’s observed score on a test or other measure (e.g. behaviour observation, opinion survey) to the average score that person would have received under all the possible conditions that the test user would be equally willing to accept” [15 p.1]. Thus, as IPC according to the PINCOM is complex, several conditions (or facets) may produce variance, in the measurement of the construct.

GT was originally introduced as a response to limitations of classical test theory (CT) which is based on the notion that each test score has a true single score, belongs to one family of parallel observations, and yields a single reliability coefficient [22]. As Shavelson, Webb and Rowley point out: “CT’s usefulness, however, depends on the researcher’s ability to estimate true-score and error variances from data. With practical application of CT, we find that error variance is not a monolithic construct; error arises from multiple sources” [23 p.922].

A brief description of a clinical observer-rated instrument for violence risk assessment may illustrate this basic principle. Bjørkly [24, 25] introduced the Scale for the Prediction of Aggressive and Dangerous Behaviour in Psychotic Patients (PAD) to assess risk of aggression pertaining to 30 possible precipitating situations, such as physical contact, limit setting, specific persons, drugs/stimulants and the like. The intention was to integrate situational and interaction components of violence as facets in risk assessment procedures. The individual patient is rated separately for the 30 precipitants (facet category 1) in relation to a future within-ward and after discharge condition (facet category 2). Furthermore, the patient’s violence potential is assessed conditional to whether the patient is in an acute or better phase of illness (facet category 3). Aggression is predicted in terms of frequency (expected likelihood of aggressive behaviour) and severity (expected seriousness of injury). This is accomplished in relation to each precipitating situation (individual vulnerability), social context (inside or outside hospital), and intrapersonal fluctuations (phase of illness). In this way a measurement model was developed to obtain graded estimates

of risk in relation to three categories of fixed facets as an alternative to a binary decision on an individual’s general and context independent risk of violence. This type of multidimensional measurement design constitutes one proponent asset of G-theory.

In line with this GT is used in our study of IPC to identify the relative impact of different sources of variance and estimating these variance components. According to GT, scores on a given test procedure will have many potential sources of variation, depending on the context in which the scores were obtained. Difficulties in the test-items may cause problems; “thus generalization from the item sample to the item universe becomes less accurate” [15 p.5]. Another inaccuracy may be due to an interaction effect between items and persons as some professionals find that some items match their experiences with IPC quite well, while other professionals do not. The phenomenon of IPC, as described by the PINCOM, involves more facets, thus also a higher number of potential error variances. This will be elaborated in the section “measurement design” under Methods.

Aims of the present study

The main aim is to assess IPC by means of GT. This is done stepwise by: 1) identifying the relative impact of different sources of variance and estimating these variance components in a G-study for the full design, and 2) investigating differences between professionals from three groups of informants (schools, primary care and specialist services), regarding the impact of the variance components on the scores.

Method

Participants

A total number of 157 questionnaires were distributed to professionals engaged in interprofessional collaboration in relation to children experiencing mental health problems, in the western part of Norway. The sample may best be considered as a convenience sample, suitable for initial testing of the newly developed questionnaire PINCOM-Q. The response rate was 86% (n=134). Nineteen percent were men and 81% female. Mean age was 46 years. The following professionals (n=134) participated in the study: teachers (n=43, 32.1%), special educators (n=17, 12.7%), psychologists (n=16, 11.9%), social workers (n=14, 10.4%), primary nurses (n=13, 9.7%), child welfare workers, n = 9, 6.7%, medical doctors (n=7, 5.2%), others (n=14, 10.4%), missing (profession not registered) (n=1, 0.7%). In Table 1, the professionals are grouped according to organizational units and level of care.

Table 1. Professionals grouped by organizational units and level of care

Level of care	Services/schools	n
Primary care (n=42)	Primary health nurse	12
	School psychology services A ¹	9
	School psychology services B	5
	School psychology services C	4
	Child protection centre	12
Specialist services (n=40)	Child psychiatric clinic D	23
	Child psychiatric clinic E	17
Schools (n=52)	School F	30
	School G	22
Total	Nine organizational units	134

¹ The capital letters A–G indicate different organizational units. Professionals may work within the same health care level, within the same service, but in different organizational units (for example in different communities).

A GT measurement design

The data reported here are part of a larger study based on a descriptive and explorative design [2, 20, 26]. The present study may also be characterized as an internal domain study, which is defined as studies that “examine relationships between observed indicators represent-

ing a defined construct” as different from external reference studies which “investigate relationships between constructs” [18, p.49]. Generalizability studies often include a large number of specific observations, and the number of items is often larger than the number of participants: “A G-study makes an explicit separation of empirical information into facets of observation and objects or targets of measurement, respectively. A facet represents the set of all acceptable conditions of observation of a particular kind...” [27]. As noted above, PINCOM and the development of PINCOM-Q, the measurement of perception of IPC involves a more complex design including several facets, as illustrated in Figure 2.

Measuring IPC within a single facet design could be described, in GT terms, as a $p \times i$ design, where p = persons and i =items. A one-facet design is defined by one source of measurement error, that is, by a single facet [15]. A uni-facet design implies that persons and items are crossed and persons receive the same kind of items. Measurements may be much more complex than this, as measurement errors may arise due to multiple sources of influence on the test scores. The present study could be expressed as: $(p:g) \times (i:t:l) \times (c)$, which is a so-called crossed and nested four-

CONTEXT (c)	BETWEEN ORGANIZATIONS						WITHIN ORGANIZATIONS					
LEVELS (l)	INDIVIDUAL		GROUP		ORGANIZATIONAL		INDIVIDUAL		GROUP		ORGANIZATIONAL	
THEMES (t)	C1 C2 C3 C4	C5 C6 C7 C8	C9 C10 C11 C12	C1 C2 C3 C4	C5 C6 C7 C8	C9 C10 C11 C12	C1 C2 C3 C4	C5 C6 C7 C8	C9 C10 C11 C12	C1 C2 C3 C4	C5 C6 C7 C8	C9 C10 C11 C12
ITEMS (i) [*]	1.....16		17.....32		33.....48		1.....16		17.....32		33.....48	
PERSONS (p:g) P1 SCHOOL P52												
P53 PRIMARY CARE P94												
P95 SPECIALIST SERVICES P134												

*Items: Four items are categorized within each of the themes (C1–C12); C1 includes items 1–4, C2 includes items 5–8 etc.

Figure 2. Data collection design.

facet design. In [Figure 2](#), the persons within group ($p:g$) indicate the “objects of measurement”. In GT, observations are admissible from a universe defined by a set of facets and how these facets are organized [\[28\]](#). Hagtvet [\[28\]](#) asserts that facets “serves to emphasize the distinction between the unit of analysis or the object of measurement (e.g. persons, students, classes, groups) and the conditions of observations” (p.249). Thus the ways the facets are organized in a given study have impact on how the scores are to be interpreted. Facets in this study are contexts (internal or external collaboration), items (which are considered to be exchangeable from a universe of items), themes (which are considered as relevant for IPC) and levels (the organization of the themes (sub-constructs in PINCOM, C1–C12) on three levels individual- group- and organizational level). In GT, all relevant sources of influence on a given score are identified and estimated in a so-called G-study. A G-study estimates all identified variance components¹. In a G-study all facets are considered as random.

It should be noted that the variance components contributing to measurement error are somewhat different for relative and absolute decisions: for relative decisions, variance components that influence the relative standing of individuals contribute to error, while in absolute decisions all variance components except the object of measurement contribute to measurement error [\[15\]](#). Consequently, relative decisions are highly relevant since we are interested in how professionals perceive IPC, *relative* to the other professionals.

The questionnaire PINCOM-Q

To ensure good representation of the constructs (C1–C12), a pool of items was formulated as close to the definition of the twelve constructs in PINCOM (see [Figure 1](#)) as possible. This shows the connection between the sub-constructs in the PINCOM and the measurement domain [\[13, 18\]](#). The items were based on acknowledged definitions found in the literature for each sub-construct (themes, C1–C12) [\[2\]](#). A complete version of the PINCOM-Q has previously been published [\[2\]](#).

The items were formulated as statements and rated on a 7-point Likert scale, ranging from “strongly agree” (1) to “strongly disagree” (7). The subjects were instructed to imagine themselves in the two contextual settings (internal collaboration and external collaboration) and

¹A subsequent D-study applies the G-study variance components to estimate or suggest a set of test conditions under which the scores are generalizable (are conditions fixed or random?). A D-study may help us to design the best possible application of the measurement procedure. It should be noted that we have not used D-studies in this investigation but suggested some possible applications for future studies in the discussion.

were given the following instructions, each part containing 48 items:

Part I: Interprofessional Collaboration — Other services: Imagine yourself collaborating with professionals from other services (for example The Child Protection centre, The Child Psychiatric Clinic, Primary Nurse, School Psychology Services and so on) and state your opinion below.

Part II: Imagine yourself collaborating with professionals from your own organization. (Organizations here are: The Child Protection Center, The Child Psychiatric Clinic, Primary Nurse, School Psychology Services and the School.) State your opinion below.

This gave a total number of 96 items in the PINCOM-Q: 2 contexts (own versus other organization)×3 contextual levels (individual, group and organizational)×4 constructs per level (themes C1–C12)×4 items ([Figure 1](#)). In sum, the PINCOM-Q was filled out by 134 professionals, giving a maximum of 12,864 observations.

Data analyses

The software programs GENOVA [\[29\]](#) and urGENOVA [\[30\]](#) were used in the analysis, as these are especially designed to deal with the facet structure of GT designs. First, a G-study was performed to disentangle the different sources of variance for the full design. urGENOVA was used because the design in the present study was unbalanced, due to different number of participants in each group. This of course was not a deliberate choice, as the sample was a convenience sample of professionals working within child and adolescent mental health care. Next, three G-studies were performed for three independent groups (schools, primary care and secondary care), to identify the relative impact of different variance components on the scores.

Missing data

Of a total of 12,864 potential scores for the sample ($n=134$) on the PINCOM-Q in the present study, only a total number of 86 (7%) were missing. It is not likely that this proportion of missing values would have affected the scores substantially. This is further confirmed by the fact that the missing data were evenly distributed on item, theme, person, and group. Missing values were replaced by MEAN (and approximated to the closest integral number).

Results

The GT analysis followed two steps. Firstly, the full design ([Table 2](#)) was analysed by urGENOVA. The

relative score variance components are presented here, since our interest is primarily in the relative ranking of professionals [15]. Object of measurement was *persons within groups (p:g)*. Secondly, based on the results in step 1, an analysis of three groups of professionals was done independently, identifying the impact of different sources of variances within each group. Object of measurement was persons (*p*).

Step 1: Estimating the relative importance of variance components (G-study)

In Table 2, the estimated variance components of persons and person/facet interaction are presented.

Table 2. Estimated G-study variance components in the (p:g)×(i:t)×(c) design

Source	Variance components	df	o ²	%	o'(o ²)
p:g	o ² pg	131	0.22644	12.10	(0.051)
pc:g	o ² pc:g	131	0.08359	4.47	(0.007)
pl:g	o ² pl:g	262	0.00814	0.43	(0.004)
pt:gl	o ² pt:gl	1179	0.17110	9.14	(0.001)
pi:gt:l	o ² pi:gt:l	4716	0.48909	26.13	(0.000)
pcl:g	o ² pcl:g	262	0.03373	1.80	(0.000)
pct:gl	o ² pct:gl	1179	0.11049	5.90	(0.000)
pci:gt:l	o ² pci:gt:l	4716	0.74914	40.02	(0.000)

a² = estimated variance component, % = % of explained variance, o'(o²) = standard error of variance component.

The *p:g* variance component (object of measurement) explains 12.1% of the variance. This indicates that there are relatively large amounts of variance describing *persons within groups* in how they perceive IPC. The interaction between persons and level within groups (*pl:g*) as well as how this interaction was modified by context (*pcl:g*) explained only a small amount of variance. The interactions between person and context (*pc:g*), person and themes (*pt:gl*) as well as the triple interaction, *pct:gl*, explained relatively substantial amount of variance; 4.47%, 9.14% and 5.90%, respectively. As expected, the complex components involving different types of inconsistencies among items (*pi:gt:l*) and (*pci:gt:l*) explained the largest amount of variance. This was expected because the G-study variance components estimate the relative importance of variance in an average item in the universe of admissible observations. These components would serve as different estimates of measurement errors in a D-study estimation.

Step 2: Estimating the relative importance of variance components in different groups (G-study)

Because the variance components presented in Table 2 were nested within groups, differences among groups were hidden. In Table 3 estimated variance components for the three groups in this study, school, primary — and secondary care, are presented.

Table 3. Estimated G-study variance components for schools, primary care and specialist services

	Source	df	VC	90 % CI	% var. (rank)
School	p	49	0.16372	0.0726–0.2986	7.7 (4)
	pc	49	0.15700	0.1026–0.2470	7.4 (5)
	pl	98	0.00590	0.0000–0.0572	0.3 (8)
	pt:l	441	0.24371	0.1813–0.3145	11.5 (3)
	pi:t:l	1764	0.48586	0.4307–0.5441	22.8 (2)
	pcl	98	0.04227	0.0132–0.0820	2.0 (7)
	pct:l	441	0.15349	0.1135–0.2000	7.2 (6)
	pci:t:l	1764	0.87478	0.8284–0.9254	41.1 (1)
Primary care	p	41	0.26509	0.1717–0.4270	15.0 (3)
	pc	41	0.05088	0.0263–0.0917	2.9 (6)
	pl	82	-0.01373	0.0000–0.0207	0 (8)
	pt:l	396	0.12869	0.0814–0.1829	7.3 (4)
	pi:t:l	1476	0.54378	0.4882–0.6034	30.8 (2)
	pcl	82	0.03225	0.0115–0.0616	1.8 (7)
	pct:l	396	0.05998	0.0322–0.0923	3.4 (5)
	pci:t:l	1476	0.68369	0.6442–0.7272	38.7 (1)
Secondary care	p	39	0.26934	0.1731–0.4393	16.1 (3)
	pc	39	0.03261	0.0119–0.0659	1.9 (6)
	pl	78	0.02129	0.0000–0.0656	1.3 (8)
	pt:l	351	0.13092	0.0828–0.1859	7.8 (4)
	pi:t:l	1404	0.42019	0.3704–0.4734	25.1 (2)
	pcl	78	0.02386	0.0017–0.0551	1.4 (7)
	pct:l	351	0.10523	0.0724–0.1440	6.3 (5)
	pci:t:l	1404	0.67210	0.6323–0.7160	40.1 (1)

The results presented in Table 3 show that there are between-group differences in the size of some of the variance components. First, the group *school* has a smaller *p*-component, than the other two groups. This could indicate that professionals in schools are more homogeneous with regard to how they perceive IPC than professionals in the other groups. However, looking at the other variance components, within each group, gives rise to another explanation. The group *school* has a higher *pc*-component relative to primary and secondary care groups. This indicates that the person component in schools is more contextually dependent. One may interpret this to be that the school personnel perceive that doing IPC within the school context (internal collaboration) is quite different from working with IPC with professionals from primary care and/or secondary care (external collaboration). Furthermore, the group *school* has a higher *pt:l* component than the other two groups. This could indicate that professionals in schools differentiate IPC more than professionals in the other two groups, with regard to which themes are relevant in IPC work, and possibly also across contexts as indicated by the *pct:l* component. In sum the results from the three G-studies presented in Table 3, indicate that persons within schools are more contextually dependent in their perceptions of IPC, than professionals in the other two groups.

Discussion

The main purpose of this paper has been to present aspects of GT as a theoretical and methodological approach in assessing the construct of interprofessional collaboration (IPC). In step 1 the relative impact of different sources of variance were identified and estimated in a G-study for the full design. In step 2 three groups of informants (schools, primary care and specialist services) were investigated to see if the estimated sources of variance differed between the groups.

Main findings

In Step 1, the findings showed that the PINCOM-Q was able to differentiate among *persons within groups* (*p:g*). This reflects how professionals in some respects perceived important aspects of IPC differently depending on what group they belonged to (school, primary care and secondary care). We do not know what caused these differences, but we may speculate that differences in educational background, experience with IPC, and the like to have an impact on this. However, other sources also affected the variability of the scores. For example, the estimated variance component $\sigma^2_{pt:gl}$, showed that variation arises from the

interaction between *persons and themes within groups and levels*. This indicates that group membership to some extent also explains the perceived importance of different IPC themes.

Step 2 (Table 3) was performed to investigate if the estimated explained variance components differed between the groups. The universe score for *persons within schools* was to a higher degree caused by variability due to interaction between *person and theme within level* (*pt:l*) than person (*p*) alone which was the case in the other groups. This may indicate that persons in schools belong to a different culture and that IPC is perceived to be more complex and context dependent than reported by persons in the other groups. Apart from this difference the relative ranking of the eight sources of variance is identical for the primary and secondary group, and the school group ranking is only marginally different from the other two groups.

Our study illustrates that in contrast to test construction within the classical test theory framework, GT gives new possibilities for evaluating test scores. One is that GT highlights both validity and reliability issues. This seems to be of great importance in measurements of IPC as the construct is multifaceted. In classical test theory, the reduction of items in a given test may be done by examining (post hoc) the degree to which each item contributes to the error variance. This strategy has previously also been used in the study of IPC [2]. In GT, both the design of the study and the way the facets are defined (random or fixed) are of great importance when considering the consistency (generalizability) of the scores.

In general, the identification of different sources of estimated variance components in GT, may help us to measure perceptions of IPC in a reliable way. According to Shavelson and Webb [15], as test users we are not interested in a person's score on a specific test. Generalization means that scores are dependent, i.e. they have an "accuracy of generalizing from a person's observed score on a test or other measure to the average score that person would have received under all the possible conditions that the test user would be equally willing to accept" [15] (p.1). It follows that if several conditions of the test situation are fixed, then generalization is restricted. The limit of the universe is thereby set by the researcher who decides what *facets* are random and what facets are fixed.

Conceptual issues

Our assumption of IPC as a multifaceted phenomenon was supported and emphasized in the GT analyses. The findings in Steps 1 and 2, showed that variability in the scores emerged related to several facets and

the interaction of these. It is possible that the multifaceted nature of IPC makes the phenomenon difficult to assess and this may explain why there still are relatively few empirical studies within this field [8, 19]. As pointed out in the introduction, many authors have tried to conceptualise the meaning of IPC, and this conceptual development of IPC may foster an increase in empirical research in the near future, including the development of different measurement procedures [1]. Hopefully, this will help bringing the field closer to an adequate understanding and reliable measurement of the phenomenon.

The development of IPC measurements, will probably give rise to further debate about how IPC should be understood. In this study, and in a previous study [2] as well, it has been suggested that the main construct of IPC may be theoretically defined by twelve sub-constructs (PINCOM, C1–C12). Although the PINCOM-Q includes a number of items representing this range of sub-constructs considered to be of central importance to IPC, the selected sub-constructs also constrain how IPC is empirically investigated. Although our preliminary findings are promising in some respects they raise further questions: What other sub-constructs could have been included? Should some sub-constructs be taken out? Furthermore, it is obvious that the twelve sub-constructs in the PINCOM represent a very challenging complexity within the field of psychology and related disciplines. It is required that the tentative descriptions of each of these sub-constructs will be improved over time, as their relation and relevance to IPC develops. This could then improve the selection of items representing the constructs, as other items could replace existing items included in the measurement domain. It is quite possible that some of the items chosen to be included in the questionnaire PINCOM-Q only partly represent the domain they are supposed to represent. Thus the domain may be said to have “fuzzy-edges” [13]. It is possible that this is a general problem in the applied methodology of measurement, as Hagtvet and Zou assert: “The current trend is rather one in which measures stay unchanged over years, sometimes in spite of substantial conceptual changes recognized in a field” [18 p.51]. Thus “sharpening” the sub-constructs in the PINCOM and clarification of the empirical domains would counteract this problem. The lack of a measurement tradition within the IPC field makes this especially important.

Another problem that may hamper construct validation is the possible overlap between domains at item level. For example, how different are the items representing *communication* and *social support* in PINCOM and why some items fit both domains? This is basically a conceptual problem, and calls for a conceptual analysis of item indicators. For example, the sub-constructs

in PINCOM (theoretical level) are considered to be descriptions of different features of the IPC. Still it is reasonable that they are related as they all substantially deal with IPC. The items representing the sub-constructs may also be correlated: a) within each domain and b) between domains. Investigating whether items overlap then could be done by examining the wording of the items [18]. As the PINCOM-Q consists of a relatively large number of items; this will likely reduce the chances of losing or under-representing essential features of the main construct of interprofessional collaboration [14]. However, the development of PINCOM is at an early stage, and future studies need to address this and related issues.

Limitations and suggestions for future research

PINCOM was developed to capture core aspects of IPC within the field of delivering mental health services towards children and adolescents. GT is a rather advanced methodology that clinicians and researchers might find difficult to apply in a given study. However, GT also give the researcher possibilities to “assess the major sources of variation so that unwanted variation can be reduced in collecting future data” [15 p.6]. In the present study the complexity in investigating the perception of the IPC construct, as illustrated by the data collection design (see Figure 2), show that there are many facets in the measurement of IPC and this produces a high number of potential sources of variance. However, the exploration of IPC using GT, may over time produce a better understanding of the phenomenon. For example, problems related to domain specification, appearing in this study, may give rise to a debate of what sub-constructs to be included in IPC models and subsequently what items that should be included in IPC tests. Also, the identification of what causes variability in the scores (facets and their organization) may give interesting information about the nature of IPC.

As the PINCOM-Q seems to have the potential of detecting differences in how IPC is perceived by professionals or groups of professionals, there could be several interesting ways to use this instrument in future studies. First it could be used to investigate how professionals in a clinical context perceive IPC, for example in internal and/or external IPC processes. The results could be used to further develop collaboration among professionals, for example as a point of departure for discussions and dialogues among the professionals and between professionals and leaders. For example, some professionals may find organizational factors essential when explaining the team’s problems, while others may be focused on group factors, such as communication

problems. Identifying such differences may help professionals to further progress in the IPC process towards a common understanding of how the actual IPC group should develop and function. High quality services are strongly linked to the standard of collaboration.

Secondly PINCOM-Q could be used to evaluate changes in perception of IPC, for example investigating professionals' perception of IPC before and after organizational changes. Furthermore, using PINCOM-Q in other samples and/or in other contexts would give new possibilities of conducting G-studies to estimate variance components and an increase in the number of observations will stabilize these components.

The G-study variance estimations, presented in this study, could also be used to design a revised version of the PINCOM-Q, in a cost-benefit perspective. For example, under what conditions would the questionnaire produce acceptable levels of generalizability of the scores? In general in a D-study, the researcher a) defines the universe of generalization (the number of facets and whether these are random or fixed), b) uses the estimations derived in the G-study to evaluate which designs that obtain adequate generalizability, and c) decides if the interpretation of the measurement should be relative or absolute [15]. In the case of IPC and based on the findings in our study, relevant questions for new studies applying a revised version of the PINCOM-Q would be: Who are the informants (school personnel or persons working in health- and social care), what facets should be included and how should these be organized, how should IPC be conceptualized (cf. domain specifications) and how many items are needed to represent the individual theme?

It should be mentioned that the urGENOVA program does not accept incomplete data matrices, thus missing values must be replaced, or all data for the person must be removed. In this study we substituted missing values by mean scores. This is disputable, especially when the number of missing values is high. An alternative solution could have been to use the regression approach in the Missing Value Analysis (MVA) in SPSS or other techniques to estimate missing values [31].

The measurement of IPC is in an early phase of development and there are numerous issues to be addressed. The most important one seems to be the definitional issue. Considering domain specifications, is an important step towards reliable and valid item generation, since validation is seriously hampered if it is unclear what constructs the items actually represent. As Nunnally points out: "most measures should be kept under constant surveillance to see if they are behaving as they should" [13 p.87].

Reviewers

L.J. Middel, Dr., Methodological consultant, Department of Health Sciences, Department of Oral Health and Clinical Epidemiology, University Medical Centre Groningen (UMCG), University of Groningen, Groningen, The Netherlands

M. Spreeuwenberg, Drs, Department of Methodology, Faculty of Social Sciences, University of Tilburg, Tilburg, the Netherlands

One anonymous reviewer

References

1. Bronstein LR. Index of interdisciplinary collaboration. *Social Work Research* 2002;26(2):113–26.
2. Ødegård A. Exploring perceptions of interprofessional collaboration in child mental health care. *International Journal of Integrated Care* [serial online] 2006 Dec 18; 6. Available from: <http://www.ijic.org>.
3. Lindqvist S, Duncan A, Shepstone L, Watts F, Pearce S. Development of the "Attitudes to Health Professionals Questionnaire" (AHPQ): a measure to assess interprofessional attitudes. *Journal of Interprofessional Care* 2005;19(3):269–79.
4. Mackay S. The Role Perception Questionnaire (RPQ): a tool for assessing undergraduate students' perceptions of the role of other professions. *Journal of Interprofessional Care* 2004;18(3):289–302.
5. Hind M, Norman I, Cooper S, Gill E, Hilton R, Judd P, et al. Interprofessional perceptions of health care students. *Journal of Interprofessional Care* 2003;17(1):21–34.
6. Tunstall-Pedoe S, Rink E, Hilton S. Students attitudes to undergraduate interprofessional education. *Journal of Interprofessional Care* 2003;17(2):161–72.
7. Leathard A. *Interprofessional collaboration: from policy to practice in health and social care*. Hove and New York: Brunner-Routledge; 2003.
8. San Martin Rodriguez L, Beaulieu M-D, D'Amour D, Ferrada-Videla M. The determinants of successful collaboration: a review of theoretical and empirical studies. *Journal of Interprofessional Care* 2005;Supplement 1:132–47.
9. Zwarenstein M, Atkins J, Barr H, Hammick M, Koppel I, Reeves S. A systematic review of interprofessional education. *Journal of Interprofessional Care* 1999;13(4):417–24.
10. Kodner DL, Spreeuwenberg C. Integrated care: meaning, logic, applications, and implications – a discussion paper. *International Journal of Integrated Care* [serial online] 2002 Nov 14; 2, Available from: <http://www.ijic.org>.

11. Barr H, Koppel I, Reeves S, Hammick M, Freeth D. *Effective interprofessional education: argument, assumption & evidence*. Oxford: Blackwell; 2005.
12. Meads G, Ashcroft J. *The case for interprofessional collaboration*. Oxford: Blackwell; 2005.
13. Nunnally JC. *Psychometric theory*. New York: McGraw-Hill; 1967.
14. Messick S. Validity of psychological assessment. *American Psychologist* 1995;50(9):741–9.
15. Shavelson RJ, Webb NM. *MMS generalizability theory: a primer*. Newbury Park: Sage Publications; 1991.
16. Hinkin TR. A review of scale development practices in the study of organizations. *Journal of Management* 1995;21(5):967–88.
17. DeVellis RF. *Scale development: theory and applications*. Newbury Park: Sage Publications; 1991. (Applied social research methods series; 26).
18. Hagtvet KA, Zuo L. Conceptual and empirical components of an internal domain study: an illustration in terms of the Achievement Motives Scale. *Scandinavian Journal of Educational Research* 2000;44(1)49–78.
19. D'Amour D, Ferrada-Videla M, San Martin Rodriguez L, Beaulieu M-D. The conceptual basis for interprofessional collaboration: core concepts and theoretical frameworks. *Journal of Interprofessional Care* 2005; Supplement 1:116–31.
20. Ødegård A. Perceptions of interprofessional collaboration in relation to children with mental health problems: a pilot study. *Journal of Interprofessional Care* 2005;19(4):347–57.
21. Cronbach LJ, Gleser, GC, Nanda H, Rajaratnam N. *The dependability of behavioural measurements: theory of generalizability of scores and profiles*. New York: John Wiley; 1972.
22. Nunnally JC, Bernstein IH. *Psychometric theory*. 3rd edition. New York: McGraw Hill; 1994.
23. Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. *American Psychologist* 1989;44(6)922–32.
24. Bjørkly S. Scale for the prediction of aggression and dangerousness in psychotic patients: an introduction. *Psychological Reports* 1993 Dec;73(3 Pt 2):1363–77.
25. Bjørkly S. The scale for the prediction of aggression and dangerousness in ppsychotic patients (PAD): a prospective pilot study. *Criminal Justice and Behavior* 1994;21:341–56.
26. Ødegård A. Time used on interprofessional collaboration in child mental health care. *Journal of Interprofessional Care* 2007;21(1):45–54.
27. Hagtvet KA, Hanin YL. Consistency of performance-related emotions in elite athletes: generalizability theory applied in the IZOF model. *Psychology of Sport and Exercise* 2007 Jan;8(1):47–72. Available from: <http://www.sciencedirect.com>.
28. Hagtvet KA. The function of indicators and errors in construct measures: an application of generalizability theory. *Journal of Vocational Education Research* 1997;22(4):247–66.
29. Crick JE, Brennan RL. *Manual for GENOVA: A GENeralized Analysis Of Variance System (Version 2.1)*. Iowa City, IA: The University of Iowa; 1983.
30. Brennan RL. *Manual for urGENOVA*. Iowa City, IA: Iowa Testing Programs, The University of Iowa; 2001.
31. Maassen G. Handling of missing values in path models for opinions or attitudes. *European Journal of Social Psychology* 1996;26(11):1–13.