



UNIVERSIDAD DE GRANADA

FACULTAD DE PSICOLOGÍA

Departamento de Metodología de las Ciencias del Comportamiento

TESIS DOCTORAL

**UTILIDAD DE LOS MÉTODOS DE PRETEST COGNITIVO
PARA OPTIMIZAR LA CALIDAD DE LOS CUESTIONARIOS Y
APORTAR EVIDENCIAS DE VALIDEZ: UNA APROXIMACIÓN
DE INVESTIGACIÓN MIXTA**

USEFULNESS OF COGNITIVE PRETEST METHODS TO
OPTIMIZE THE QUALITY OF QUESTIONNAIRES AND TO
PROVIDE VALIDITY EVIDENCE: A MIXED RESEARCH
APPROACH

Isabel Benítez Baena

Director: José Luis Padilla García

Universidad de Granada

Co-directora: Juana Gómez Benito

Universidad de Barcelona

Editor: Editorial de la Universidad de Granada
Autor: Isabel Benítez Baena
D.L.: GR 3109-2012
ISBN: 978-84-9028-233-5

Utilidad de los métodos de pretest cognitivo para optimizar la calidad de los cuestionarios y aportar evidencias de validez: una aproximación de investigación mixta.

Usefulness of cognitive pretest methods to optimize the quality of questionnaires and to provide validity evidence: A Mixed Research approach

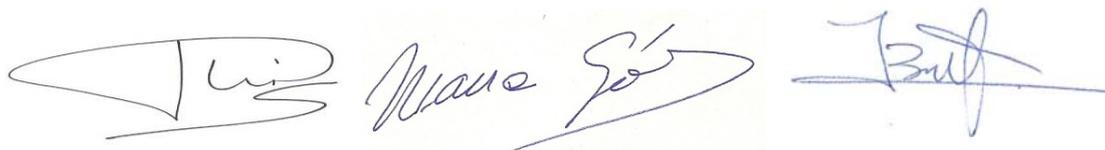
Tesis doctoral presentada por Isabel Benítez Baena en el Departamento de Metodología de las Ciencias del Comportamiento de la Universidad de Granada, dentro del Programa Oficial de Doctorado en Psicología, para aspirar al grado de “**Doctor en Psicología**” con mención de “**Doctor Internacional**”.

Doctoral dissertation presented by Isabel Benítez Baena in the Department of Methodology of Behavioural Sciences of the University of Granada (Programa Oficial de Doctorado en Psicología) to obtain the degree of “**PhD in Psychology**” with the distinction of “**Doctor Internacional**”.

La tesis ha sido dirigida bajo la dirección del doctor José Luis Padilla García y la codirección de la doctora Juana Gómez Benito, quienes avalan la calidad de la misma y la formación de la doctoranda para aspirar al grado de doctor.

The doctoral dissertation has been supervised by Dr. José-Luis Padilla García and co-supervised by Dr. Juana Gómez Benito, who are acting as guarantors for the quality of the thesis and the aspirant's qualifications to obtain the PhD degree.

En Granada, a 20 de Abril de 2012



Fdo.: José-Luis Padilla García

Director

Fdo.: Juana Gómez Benito

Co-directora

Fdo. Isabel Benítez Baena

Doctoranda

A mis padres, a mi hermana

y a Oscar,

Os llevo siempre conmigo

Agradecimientos

Me gustaría dar las gracias a las personas que me han acompañado a lo largo de este camino.

Mi mayor agradecimiento es para José Luis, por muchos motivos, pero principalmente por haber tenido desde el primer día confianza ciega en mí y por velar siempre por mis intereses. Gracias por ser un amigo además de “mi jefe”.

A mi familia, a mi madre y a mi hermana, por estar ahí cada día. A mi padre, por inculcarme valores como la constancia y la pasión por el trabajo que tan importantes han sido durante este proceso.

A Miguel por ser mi compañero, mi amigo y mi consejero, y sobre todo por los buenos momentos que hemos pasado juntos.

A Pablito por su ayuda infinita, su paciencia, y por escucharme cada día.

A “mis niñas”, porque no se me ocurre mejor vía de escape que su compañía.

A los compañeros del departamento, por asesorarme y contar conmigo. En especial a la gente del grupo, por nuestras discusiones metodológicas. A Andrés, porque siempre tiene respuestas y está dispuesto a compartirlas. A Cristino, por sus correos poéticos que nos hacen reflexionar. A José María, por ser capaz de ponernos a todos en orden. A Hugo, por la alegría que transmite. Y a Macarena, por nuestras reuniones “emocionales” y por estar siempre dispuesta a ayudar.

A las personas con las que he tenido la oportunidad de trabajar durante las estancias. A Juana, por confiar siempre en mi criterio y en mi capacidad. A Lola, por enseñarme que se puede luchar contra todo. A Yfke, por abrirme las puertas sin dudar. A Steve por motivarme y darme tanta energía, y porque aún es mejor persona que investigador. Ha sido un grandísimo placer trabajar con todos vosotros.

A las personas que he ido encontrando en las estancias y que han hecho que me sienta como en casa. A la gente de de la Universidad de Barcelona, en especial a la Jor y a la Pilar, por compartir conmigo esos cafés a primera hora. A los amigos que dejé en Amherst, especialmente a Joyce, por convertirse en tan poco tiempo en una persona tan importante para mí.

A las personas con las que he convivido durante mis múltiples estancias y que aún siguen estando presentes en mi vida. Sobre todo a Laia, porque me encantan esos momentos de “ponernos al día”.

A mis compañeros “psicólogos” porque aunque pasen los años aún nos encantamos. A Moni y a Vero, por nuestros encuentros llenos de risas. A Maryem, por nuestras conversaciones, porque aunque sean pocas siempre son intensas. A Marina y a Julián, por mantener la ilusión del reencuentro.

A todas las personas que han estado a mi lado, sobre todo en estos últimos tiempos que tan intensos han sido.

Y por supuesto a ti, Oscar, por dar luz a mi vida.

Gracias a todos por estar ahí.

Esta investigación ha sido realizada gracias a la beca del Programa Nacional de Formación de Profesorado Universitario (FPU; referencia AP2007-03118) concedida a Dña. Isabel Benítez Baena. Parte de la investigación desarrollada ha sido realizada en el marco de los proyectos de investigación concedidos por la Consejería de Innovación y Ciencia de la Junta de Andalucía al Dr. José Luis Padilla García (SEJ-5188 y SEJ-6569), y del proyecto concedido por el Ministerio de Ciencia e Innovación (PSI2009-07280).

Índice de Contenidos / Table of Contents

RESUMEN	9
SUMMARY	15
INTRODUCCIÓN.....	19
El papel de las evidencias de validez.....	22
La investigación mixta o “Mixed Research”	25
Los métodos de pretest cognitivo.....	30
Entrevistas cognitivas.....	32
Codificación del comportamiento.....	37
Los informantes indirectos o proxies.....	41
Los análisis psicométricos y el Funcionamiento Diferencial de los Ítem	45
OBJETIVOS Y ESTRUCTURA	49
INTRODUCTION.....	55
The role of validity evidence.....	58
Mixed Research	61
Cognitive pretest methods.....	64
Cognitive Interviews.....	67
Behavior coding.....	71
Proxy respondents	74
Psychometric analysis and Differential Item Functioning	78
OBJECTIVES AND STRUCTURE	81
CHAPTER 1:.....	87
Analysis of Quality of Proxy Questions in Health Surveys by Behavior	
Coding.....	87
Abstract	88
Introduction.....	89
Method	91
Participants	91

Materials.....	92
Procedure	94
Verbal behavior coding.....	94
Results	96
Types of sequence	97
Codes for proxy responses	98
Difficulty indicators when asking questions	101
Discussion.....	102
References	106
CHAPTER 2:.....	109
Evaluation of the convergence between "self-reporters" and "proxies" in a disability questionnaire by means of Behavior Coding method	109
Abstract	110
Introduction.....	111
Method	114
Participants	114
Materials.....	115
Procedure	117
Analysis	117
Results	119
Sequences types analysis	119
Convergence evaluation	122
Discussion.....	125
References	130
CHAPTER 3:.....	133
Obtaining validity evidence by cognitive interviewing to interpret psychometric results	133
Abstract	134
Introduction.....	135
The Cognitive Interviewing Reporting Framework (CIRF)	136

Overall objectives	136
Study 1: The analysis of the psychometric characteristics of the items included in the APGAR scale.....	137
Method	138
Participants	138
Materials.....	138
Data collection	140
Data analysis.....	140
Results	140
Conclusions	143
Study 2: Use of cognitive interviews to gather evidence about the answer process carried out by the respondents.....	143
Research objectives	143
Method	144
Participant selection.....	144
Materials.....	144
Research Design	144
Ethics and data collection	145
Data analysis.....	145
Findings.....	145
Discussion and Conclusions	152
Reflections on the use of the CIRF	155
References	157
CHAPTER 4:.....	161
A two method-two effect size measure strategy for analysing polytomous Differential Item Functioning: An illustration with Differential Step Functioning and Ordinal Logistic Regression	161
Abstract	162
Introduction.....	163
Statistical procedures: Differential Step Functioning and Ordinal Logistic Regression	166
Method	168

Participants	168
Instruments	169
The 2006 PISA database	170
Analyses	170
Results	171
Descriptive Statistics.....	171
DIF results: Differential Step Functioning.....	173
DIF results: Convergence across methods	177
DSF analyses for DIF items.....	178
Discussion.....	179
References	182
CHAPTER 5:.....	187
Analysis of the causes of DIF using a Mixed Methods approach: Analysis of Cognitive Interviews and DIF.	187
Abstract	188
Introduction.....	189
Quantitative phase.....	193
Participants	193
Analyses	194
Qualitative phase	195
Participants	195
Materials.....	196
Analyses	196
Mixed Research framework design	198
Results	199
Quantitative findings.....	199
Qualitative finding.....	201
Findings integration	205
Discussion.....	207
References	211
DISCUSIÓN.....	217

Utilidad de los métodos de pretest cognitivo para resolver problemas de investigación.....	219
Capacidad de los métodos de pretest cognitivo para aportar evidencias de validez basadas en los procesos de respuesta	221
¿Cuáles son los beneficios de combinar resultados siguiendo los fundamentos de la investigación mixta?	223
CONCLUSIONES Y LINEAS FUTURAS.....	225
Conclusiones finales	226
Limitaciones y líneas futuras.....	227
DISCUSSION.....	231
Use of cognitive pretest methods to resolve research problems.....	233
Capacity of the cognitive assessment methods to provide validity evidence based on the response process	235
What are the benefits of combining results following the guidelines of Mixed Research?.....	236
CONCLUSIONS AND FUTURE STUDIES	239
Final Conclusions.....	240
Limitations and future studies.....	241
REFERENCIAS / REFERENCES.....	245
ANEXOS / APPENDIX.....	253

Índice de Tablas / Index of Tables

INTRODUCCIÓN

Tabla 1. Tipos y ejemplos de pruebas cognitivas.....	34
Tabla 2. Categorías para la clasificación de los comportamientos ocurridos durante la interacción entrevistador-entrevistado	38

INTRODUCTION

Table 1. Types and examples of follow-up probe 68

Table 2. Categories for the classification of the behaviors occurring during interviewer-respondent interaction 72

CHAPTER 1

Table 1. Selected questions from the disability questionnaire. 93

Table 2. Categories for the classification of respondents' behaviors. 95

Table 3. Frequencies and percentages of each type of sequence. 97

Table 4. Frequencies and percentages of answer category codes. 99

Table 5. Frequencies and percentages of difficulty indicator codes during the question reading. 101

CHAPTER 2

Table 1. Description of characteristics of the cognitive pretest participants.... 115

Table 2. Self-reporter target question..... 116

Table 3. Responses categories to classify respondents' answers. 118

Table 4. Percentage of different sequences produced by "proxies" and self-reporters 119

Table 5. Percentages of problematic and non problematic answers 121

Table 6. Percentages of the behaviors occurring while asking the question.... 122

Table 7. Percentage of sequence with agreement and disagreement..... 124

Table 8. Agreement between proxy and self-reporter answers 125

CHAPTER 3

Table 1. Demographic characteristics of participants 138

Table 2. Original APGAR items..... 139

Table 3. Descriptive statistics for the APGAR items. 140

Table 4. Exploratory factor analysis results. 141

Table 5. Factor loadings and communalities of the APGAR items 142

Table 6. Themes developed concerning providing help.....	149
Table 7. Themes developed concerning making decisions.	151
CHAPTER 4	
Table 1. PISA scales used in the DIF analyses.....	169
Table 2. Descriptive statistics for the PISA scale items	172
Table 3. Summary of DSF Results	174
Table 4. Summary of OLR results	176
Table 5. Convergence across methods.....	177
Table 6. DSF steps, magnitude and sign in items with DIF.....	178
CHAPTER 5	
Table 1. Items with Large DIF.....	199
Table 2. Summary of DSF form results.....	200
Table 4. Finding about differences in the interpretation of “Advances in broad science and technology”	202
Table 5. Examples of narratives in item Gen 5.	203
Table 6. Examples of the “university-job” sub-theme.....	204

Índice de Figuras / Index of Figures

INTRODUCCIÓN

Figura 1. Representación modelo pregunta-y-respuesta.....	31
Figura 2. Representación gráfica del modelo piramidal	36

OBJETIVOS Y ESTRUCTURA

Figura 1. Relación entre los estudios, el método cognitivo aplicado, el diseño de investigación y la forma de combinar los resultados.....	53
--	----

INTRODUCTION

Figure 1. Representation of the question-and-answer model.....	65
--	----

Figure 2. Graphical representation of the pyramid model..... 70

OBJECTIVES AND STRUCTURE

Figure 1. Relationship between the studies, the cognitive method applied, the research design and the method of combining the results..... 85

CHAPTER 5

Figure 1. Phases of the (QUAN+QUAL) study..... 198

RESUMEN

El estudio de la calidad de las mediciones aportadas por los instrumentos de evaluación ha tenido una gran relevancia en el contexto de la investigación por encuestas y de la evaluación psicológica. Desde ambas perspectivas se ha planteado la necesidad de desarrollar procedimientos que optimicen la elaboración de tests y cuestionarios, y aumenten la calidad de la información que aportan.

El desarrollo de estos procedimientos, focalizados en incrementar la calidad de la información, ha estado ligado al avance de las necesidades emergentes en el campo aplicado, que a su vez han estado marcadas por la evolución del propio concepto de *validez*. La actual concepción de la validez señala entre las fuentes de evidencias las basadas en el proceso de respuesta de los participantes. Los *Standards* (APA, AERA y NCME, 1999), especifican que este tipo de evidencias debe extraerse del análisis de respuestas individuales, lo que señala los métodos de pretest cognitivo como una alternativa adecuada. Por otra parte, los *Standards* también aluden a la necesidad de acumular evidencias tanto cuantitativas como cualitativas, lo que refuerza la tendencia actual a la combinación de procedimientos recogida en los fundamentos de la investigación mixta (o *Mixed Research*). La investigación mixta plantea un esquema para abordar el estudio de los problemas que surgen en el campo de la investigación, mediante la combinación de procedimientos cuantitativos y cualitativos con el fin de que dicha combinación nos acerque a una mejor respuesta.

Por ello, en esta tesis se han aplicado diseños propios de la investigación mixta para abordar problemas aplicados como, por ejemplo, la evaluación de la calidad de las respuestas de los informantes proxy (personas que responden a las preguntas pensando en la situación de otra persona), la utilidad de las Entrevistas Cognitivas (EC) para interpretar resultados psicométricos y procedentes del análisis del Funcionamiento Diferencial de los Ítems (DIF), o la búsqueda de las causas del DIF.

El objetivo general de la tesis es plantear la evaluación de la calidad de las mediciones aportadas por escalas y cuestionarios, mediante diseños de investigación mixtos en los que los métodos de pretest cognitivo y los métodos psicométricos se combinen con el fin de obtener evidencias de validez basadas en los procesos de respuesta. Para alcanzar este objetivo se han planteado cinco estudios, cuyas características y resultados principales se describen a continuación.

El objetivo del Estudio 1 fue mostrar la utilización de la Codificación del Comportamiento (CC) para evaluar la calidad de las respuestas proporcionadas por informantes proxies. Para ello, se administró un cuestionario a personas que convivían con discapacitados y que respondieron sobre las limitaciones que estos tenían. Las respuestas de los informantes proxies fueron analizadas con el fin de localizar aspectos problemáticos en las preguntas a partir de las dificultades de los participantes durante la interacción entrevistador-entrevistado. Los resultados mostraron la utilidad de la CC para analizar la calidad de las respuestas ofrecidas por los informantes proxies, lo que resulta relevante en contextos en los que no siempre es posible contactar con los informantes directos y en los que el uso de los proxies ha sido cuestionado.

El Estudio 2 surge ante las conclusiones del Estudio 1, cuando se observa que la evaluación de la calidad de las respuestas de los informantes proxies se realiza a partir de la evaluación de la coincidencia entre éstas y las respuestas de los informantes directos, que son consideradas el "gold standard". El objetivo del estudio fue evaluar la convergencia de las respuestas de informantes directos y proxies a un cuestionario de discapacidad con el fin de valorar las aportaciones de la CC para realizar esta evaluación y ampliar los resultados habituales dando detalles sobre los tipos de desacuerdo.

Los resultados señalaron la CC como un método adecuado para evaluar la convergencia entre las respuestas de los informantes, por proporcionar información relevante sobre las características y los motivos del desacuerdo

entre informantes. Además se observó una alta convergencia entre las respuestas de los informantes, lograndose altos porcentajes de acuerdo.

En el Estudio 3 el objetivo era evaluar la conexión entre los resultados proporcionados por los métodos psicométricos tradicionales y por los métodos de pretest cognitivo. Se utilizaron las respuestas de los participantes a una escala de función familiar para calcular los estadísticos psicométricos habituales, y las respuestas a las pruebas de indagación de las EC para obtener evidencias de los procesos de respuesta realizados para responder. Los resultados mostraron una relación entre ambos tipos de resultados que variaron en la misma dirección, es decir, los datos estadísticos reflejaron las diferencias en las interpretaciones de los participantes. Además se demostró la capacidad de las EC para interpretar resultados psicométricos.

En el Estudio 4, se muestra una aplicación de análisis del DIF a los ítems politómicos de escalas actitudinales incluidas en el estudio PISA (Program for International Student Assessment; OECD, 2006). El objetivo del estudio fue proponer una estrategia para la evaluación del DIF basada en la aplicación de dos métodos y dos medidas del tamaño del efecto. Los resultados permitieron seleccionar con un alto de nivel de seguridad los ítems que posteriormente fueron incluidos para desarrollar el Estudio 5.

El Estudio 5 se basó en los ítems seleccionados y en la clasificación del tipo de DIF realizada en el estudio anterior. La información sobre las interpretaciones de los participantes se utilizó para comparar los procesos de respuesta entre los participantes de los distintos grupos. A continuación, se conectaron los resultados del DIF con los resultados de las EC con el objetivo de alcanzar una conclusión conjunta sobre las causas del DIF. Los datos mostraron una relación entre los ítems detectados y los procesos de respuesta, que evidenciaron diferencias a la hora de interpretar conceptos y al enmarcar experiencias en un contexto.

Las conclusiones formuladas a partir de la integración de estos resultados fueron las siguientes:

- a) Los métodos de pretest cognitivo son útiles para resolver problemas metodológicos más complejos que los que habitualmente se abordan en la evaluación de los cuestionarios de encuestas .
- b) Los métodos de pretest cognitivo aportan evidencias de validez valiosas sobre los procesos de respuesta de las personas que responden a los cuestionarios y escalas.
- c) Utilizar distintos procedimientos dentro de un paradigma de investigación mixta facilita la obtención de conclusiones más ricas que las proporcionadas por los métodos de forma separada.

SUMMARY

Ensuring the quality of information provided by assessment instruments has been relevant in the survey research and psychological testing contexts. Both coincide in the growing interest in developing procedures to optimize test and questionnaire elaboration and to improve the quality of the information provided.

Development in procedures and methods focused on increasing the quality of the information has been related to changes in emerging needs in the applied field which have been affected by the evolution of the concept of validity itself. Evidence based on response processes are one of the sources of validity pointed out in the current concept of validity. The *Standards* (APA, AERA y NCME, 1999), specify they be drawn from the analysis of individual responses, so cognitive pretest methods could be an adequate option.

On the other hand, *Standards* also refer to the need of accumulate both quantitative and qualitative evidence, reinforcing the new trends that combine different types of methodologies included in the Mixed Research (MR) principles. MR proposes a scheme for studying topics in the field context through the combined use of quantitative and qualitative methodologies for establishing better conclusions.

It is the reason mixed designs have been applied in this thesis for resolving applied problems as the quality of proxy (people who respond by thinking in self-reporter situation) responses; the utility of Cognitive Interviews (CI) for interpreting psychometrics and Differential Items Functioning (DIF) results; or the searching of sources of DIF.

The overall objective of the thesis was to present an evaluation of the quality of measurements provided by scales and surveys, using a MR design in which cognitive pretest methods as well as psychometric methods were combined to provide validity evidence based on respondents' response processes. Five studies have been developed for responding to this objective, whose main characteristics and results are described.

The aim of Study 1 was to show how to analyze the quality of proxy informants' responses by means of Behavior Coding (BC). To do it, people who lived with people with limitations responded to a disability questionnaire about self-reporter situation. Proxy responses were analyzed for locating problematic questions by detecting participants' difficulties during interviewer-respondent interaction. Results confirm the utility of BC for analyzing quality of proxy informants responses, which is especially relevant in contexts in which self-reporter cannot be contacted and the proxy utilization has been doubted.

Study 2 came up conclusions from Study 1, when it was detected the quality of proxy responses is evaluated by comparing it with self-reporter responses, which are considered the "gold standard". The aim of this study was to evaluate the convergence between self-reporter and proxies answers to a disability questionnaire, for knowing the BC contributions to the convergence evaluation and to increase the habitual results by giving details of the disagreement type. Results indicate BC is an appropriate method for evaluating the convergence between responses, because it provides relevant information about the characteristics and causes of the disagreement across informants. Also, it was observed a high convergence between the informants' responses, reaching high percentages of agreement.

In the Study 3 the aim was to evaluate the connection between results provided by psychometrics analyses and cognitive pretest methods. Responses to a family function scale were used for calculating habitual psychometric statistics, and responses to follow-up probes in a CI protocol were used for obtaining evidences based on the respondents' response processes. Results showed a relation between both types of results that varied in the same direction, that is, statistics results shows the utility of CI for interpreting the results from psychometric analysis.

Study 4 illustrated an application of DIF analyses to the items of attitudinal scales included in the Program for International Student Assessment (PISA; OECD, 2006). The aim of the study was to propose a two method-two effect size

measure strategy for analyzing polytomous DIF. Results allowed to increase confidence in DIF items selected which were used for developing Study 5.

Lastly, Study 5 was based in items selected and in the DIF classification in the previous study. Information about participants' interpretations was used for comparing response processes of participants from different groups. Later, DIF results were connected with evidences from CI for reaching a common conclusion about DIF causes. Results showed a relation between items flagged and the response processes which indicate differences when interpreting concepts and when enshrined experiences in a specific context.

Conclusions from results' integration were:

- a) The cognitive pretest methods are useful for resolving more uncommon complex methodological problems that were found in routine assessment survey questionnaires.
- b) Cognitive pretest methods provide valuable validity evidence based on respondents' response processes to questionnaires and scales.
- c) The use of different procedures within a MR paradigm allows firmer conclusions to be drawn than those provided by individual methods.

INTRODUCCIÓN

Los instrumentos de evaluación psicológicos y sociológicos, actualmente, forman parte de nuestra vida diaria. Nos hacen encuestas de satisfacción las compañías telefónicas; en los hoteles evalúan nuestra experiencia; en los aeropuertos, nuestros hábitos de viaje y en los centros educativos, nuestro rendimiento. Sin darnos cuenta respondemos a las preguntas de manera cada vez más automática y no percibimos la importancia y las implicaciones de dichas evaluaciones. Sin embargo, detrás de todos esos instrumentos, existe un proceso largo y complejo cuyo objetivo final es asegurar que la información obtenida es de calidad y que responde a las necesidades planteadas. Asegurar la calidad de la información que se obtiene obliga a disponer de metodologías que permitan evaluar la idoneidad de las mediciones en distintos ámbitos como son las encuestas de satisfacción, los cuestionarios para la evaluación psicológica, los tests de rendimiento, los estudios transculturales, etc.

La búsqueda de metodologías que mejoren la calidad de las mediciones se ha desarrollado separadamente en dos contextos: el contexto de la investigación por encuesta y el de la evaluación psicológica. Ambos contextos han sido históricamente diferentes en aspectos tan relevantes como la atención prestada a los diferentes tipos de error (Groves, 1989; Van de Vijver, 1998). Sin embargo, coinciden actualmente en la insatisfacción por el tratamiento habitual de la calidad de las mediciones, mostrando un interés creciente por el desarrollo de procedimientos para optimizar la elaboración de tests y cuestionarios y aumentar la calidad de la información que aportan.

En el contexto de la investigación mediante encuestas este interés se refleja en la renovada preocupación por los errores de medida, que junto con los errores de muestreo, los errores de cobertura y los errores de no respuesta forman la clasificación tradicional de las fuentes principales de error propuesta por Groves (1989). Los errores de medida incluyen los llamados “errores de observación” que recogen fuentes de sesgo relacionadas con el instrumento de medida, el entrevistador, el entrevistado y el método de recogida de datos

(Groves et al., 2004). Estas fuentes de sesgo reflejan la ampliación de los puntos de interés que ahora van más allá del contenido del instrumento.

Paralelamente, en el contexto de la evaluación psicológica, el análisis de la calidad de las mediciones se ha realizado mediante la aplicación de procedimientos psicométricos que se han centrado en el análisis de la distribución de las respuestas a los ítems o en la capacidad de los mismos para discriminar entre las personas con distinto nivel de la variable. No obstante, con el paso del tiempo, las necesidades emergentes en el campo aplicado han ido cambiando y este cambio ha marcado la evolución del propio concepto de *validez*. Por ejemplo, la necesidad de comparar diferentes grupos demográficos, lingüísticos, culturales, etc., ha estimulado el diseño de estudios comparativos, lo que a su vez ha implicado un mayor desarrollo de los procedimientos cuantitativos y cualitativos utilizados para obtener resultados sobre las posibles diferencias entre los grupos. A su vez, el interés por mejorar la calidad de los estudios por encuesta ha incrementado la atención prestada a los métodos y por tanto, ha motivado su desarrollo y la rápida evolución de procedimientos como los métodos de pretest cognitivo. También, en este círculo de inquietudes, han tenido cabida nuevas tendencias como el diseño de investigaciones mixtas (o *Mixed Research*) que propone la combinación de procedimientos cuantitativos y cualitativos. Lo que sí es claro es que el objetivo presente, en esta evolución continua de metodologías y necesidades de investigación, es mejorar el proceso de obtención de evidencias de validez que favorezcan la formulación de conclusiones adecuadas sobre las diferencias entre los grupos.

La relación bidireccional que existe entre el desarrollo de la metodología y las inquietudes en el campo aplicado, así como la constante presencia de la validez como mediadora de las relaciones entre ambos, es también un rasgo definitorio de los estudios de esta tesis. Por ello, antes de presentar los trabajos realizados, se describirán algunos puntos relevantes que permitirán una mejor ubicación de sus contenidos. En primer lugar se presentarán los contenidos relacionados con la Teoría de la Validez, la investigación mixta o *Mixed Research*

y los métodos de pretest cognitivo. A continuación, se introducen los problemas aplicados objeto de las investigaciones como son: la utilización de los informantes indirectos o proxies y la interpretación de análisis psicométricos y del Funcionamiento Diferencial de los ítems (DIF). Por último, se expondrán los objetivos de esta tesis y se describirán los estudios teniendo en cuenta algunas dimensiones referentes a sus características metodológicas.

El papel de las evidencias de validez

Los *Standards*, nombre con el que se conoce al manual elaborado desde 1954 por la *American Psychological Association*, la *American Educational Research Association*, y el *National Council of Measurement in Education*, han articulado a través de las sucesivas ediciones, el consenso sobre la teoría y práctica en el uso de los test, siendo un referente para seguir la evolución de las concepciones sobre la validez en los distintos momentos históricos. Sin embargo, este consenso no ha estado ni está exento de debate como queda reflejado en las monografías y artículos más recientes sobre la Teoría de la Validez (e.g. Sireci, 2009; Zumbo, 2009).

La validez, ha pasado de ser un requisito examinado al final de la elaboración del test o cuestionario, a impregnar todo el proceso de evaluación, al tiempo que también han aumentado las formas y procedimientos para realizar los estudios de validación. Dos de las aportaciones más recientes y llamativas han sido la sustitución de las categorías o tipos de validez por la noción de “evidencias de validez” y “fuentes de evidencias de validez”, junto a la aparición de la aproximación a la validez basada en argumentos para orientar los procesos de validación propuesta por Kane (1992).

Centrándonos en las dos últimas ediciones de los *Standards* se pueden destacar los argumentos más relevantes del marco actual de la validez. En la cuarta edición de los *Standards* (APA, AERA y NCME, 1985) ya se reflejaba la

necesidad de concebir la validez como un concepto unitario, lejos de la ampliamente criticada división anterior en “categorías de validez”. En ese momento se abandonaron las tipologías de validez para empezar a hablar de estrategias de validación cuyo objetivo es recoger diferentes tipos de evidencias. Sin embargo, este cambio no fue suficiente ya que los investigadores no estaban satisfechos con la visión sobre la “validez de constructo” y demandaban nuevos cambios en la definición que situaran el constructo en una posición central, y que además incluyeran las consecuencias sociales del uso de los tests. Estas ideas se incorporaron en la quinta edición de los *Standards* (APA, AERA y NCME, 1999), momento en que la visibilidad de la validez, según explican Hambleton y Pitoniak (2002), había aumentado por el creciente uso de los tests en la toma de decisiones críticas para las personas e instituciones: contratación, selección, diagnóstico, graduación, etc. La quinta edición de los *Standards*, versión de referencia actualmente hasta la publicación de la sexta edición prevista para finales de 2012 o principios de 2013, extiende la validez a todas las fases del proceso de construcción y uso del test, lo que conlleva definir la validez como “el grado en el que la evidencia y la teoría apoyan las interpretaciones de las puntuaciones implicadas por los usos previstos de los tests. El proceso de validación implica acumular evidencia que proporcione una base científica sólida para las interpretaciones propuestas de las puntuaciones. Son las interpretaciones de las puntuaciones en los tests requeridas por los usos propuestos las que son evaluadas, no el test en sí.” (APA, AERA y NCME, 1999, p. 9). Como se destaca en la definición, la clave del proceso es acumular evidencia científica suficiente para apoyar la interpretación propuesta, pudiendo obtenerse esta evidencia a partir de diferentes fuentes.

Por primera vez, el concepto de fuentes de evidencias desempeña un papel clave en el proceso de validación. Los *Standards* desarrollan este concepto agrupando las fuentes de validez en cinco categorías:

- a) Evidencias basadas en el contenido del test: que proceden del análisis de las relaciones entre el contenido del test y el constructo que se pretende medir.
- b) Evidencias basadas en la estructura interna: que indican el grado en que las relaciones entre los ítems del test y los componentes del test se ajustan al constructo sobre el cual se basan las interpretaciones de las puntuaciones del test.
- c) Evidencias basadas en la relación con otras variables: que incluyen los análisis de las relaciones entre las puntuaciones en el test y variables “externas” al test. Por ejemplo, medidas de algún criterio que se espera prediga las puntuaciones del test, así como relaciones con otros tests que miden el mismo constructo, y con tests que miden constructos diferentes o relacionados.
- d) Evidencias sobre las consecuencias del uso del test: así como la evaluación de la adecuación del uso del test (Shepard, 1997), las implicaciones de valor asociadas a las interpretaciones de las puntuaciones y las consecuencias sociales asociadas al uso del test (Messick, 1989).
- e) Evidencias basadas en los procesos de respuesta: recurriendo al análisis teórico y empírico de los procesos de respuesta de las personas que responden al test, para obtener evidencias sobre el ajuste entre el constructo y la naturaleza detallada de la ejecución o respuesta realmente puesta en práctica por los examinados.

Las metodologías empleadas también deben responder a esta agrupación, es decir, están determinadas por la fuente concreta de evidencias de validez que quiera abordarse. En relación a las evidencias basadas en los procesos de respuesta, los *Standards* especifican que deben extraerse del análisis de respuestas individuales, de forma que se pregunte a las personas que responden, sobre sus estrategias para responder a las preguntas o sobre sus respuestas a los ítems, para obtener evidencias que enriquezcan la definición del constructo.

Sin embargo, la falta de indicaciones sobre cómo obtener las evidencias basadas en los procesos de respuestas y sobre cómo interpretarlas, provocó dudas en los investigadores que no sabían cómo abordar el estudio de este tipo de evidencias. Actualmente, como indican Zumbo y Shear (2011), se observa un notable crecimiento de los estudios centrados en obtener evidencias de validez basadas en los procesos de respuesta, en comparación con el número de estudios desarrollados en base a las fuentes de evidencias más tradicionales. En muchos de estos estudios, la obtención de evidencias se realiza mediante la aplicación de Entrevistas Cognitivas (EC) u otros métodos similares.

Los *Standards*, al hacer referencia a las características de las evidencias, aluden a la acumulación de evidencias, tanto cuantitativas como cualitativas, reforzando las nuevas tendencias que combinan distintos tipos de metodologías para incrementar las fortalezas de cada una y reducir las debilidades que surgen cuando se utilizan de forma exclusiva. Por ello, la investigación mixta o *Mixed Research* se presenta como un posible marco para abordar la búsqueda de evidencias de validez mediante la utilización conjunta de metodologías cuantitativas y cualitativas. A continuación, se describen los fundamentos de la investigación mixta que están presentes en todos los estudios recogidos en esta tesis.

La investigación mixta o “Mixed Research”

La investigación mixta se fundamenta en la inquietud por responder a problemas de investigación que, por su complejidad, son insatisfactoriamente abordados mediante un único método o un único tipo de datos. Las limitaciones que supone la “exclusividad” metodológica hacen que se plantee lo que se denomina “pragmatismo”, visión que prioriza la importancia del objetivo de investigación poniendo los métodos al servicio de éste. Johnson y Christensen (2008) situaron el concepto de pragmatismo en el eje central de esta

corriente por ser el que determina que lo relevante no es si los métodos se consideran cuantitativos o cualitativos, sino si facilitan el alcance de los objetivos que se persiguen.

Por tanto, lo que promulga la investigación mixta es la combinación de procedimientos cuantitativos y cualitativos en los casos en que dicha combinación nos acerque a una mejor respuesta para el problema de investigación. La mayoría de las definiciones que se han formulado de la investigación mixta hacen referencia a esta combinación de procedimientos aunque desde distintos enfoques. Una de las definiciones más relevantes podría ser la enunciada por Tashakkori y Creswell en 2007 en el primer artículo del primer número de la revista *Journal of Mixed Methods Research*. Estos autores definen la investigación mixta como “aquella investigación en la que se recogen y analizan datos, se integran hallazgos y se formulan inferencias utilizando aproximaciones o métodos cuantitativos y cualitativos en un mismo estudio o programa de investigación” (p. 4).

La investigación mixta se presenta por tanto como un nuevo paradigma que pretende conciliar el tradicional debate entre los investigadores cuantitativos y cualitativos (Reichardt y Rallis, 1994). Numerosos estudios desarrollados durante los últimos años han implementados diseños mixtos, sin embargo, autores como Johnson y Christensen (2008), consideran que su legitimización como “tercer paradigma” no llegó hasta la publicación del *Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori y Teddlie, 2003). Estos autores publicaron su primer libro en 1998 donde describieron la investigación mixta como un paradigma que aúna métodos cuantitativos y cualitativos para obtener datos más informativos y sofisticados (Tashakkori y Teddlie, 1998).

La influencia de la investigación mixta ha ido creciendo sobre todo desde la década de los 90, aunque su inicio fue anterior como muestra la revisión que realizaron Greene, Caracelli y Graham (1989). Los autores clasificaron los diseños más frecuentes en función de las características y los objetivos de los estudios que aplicaban investigación mixta. Como se ha comentado

anteriormente, en 1998 Tashakkori y Teddlie publicaron su primer libro en este área "*Mixed Methodology: Combining the qualitative and quantitative approaches*", al que siguieron otras publicaciones como el libro "*Handbook of Mixed Methods in Social and Behavioral Research*" en 2003 motivado por el creciente número de tesis que nacieron en el campo. La segunda edición de este libro vio la luz en 2010. Por otra parte, el aumento de los estudios que aplicaban los fundamentos de la investigación mixta provocó en 2007 la creación de la revista: *Journal of Mixed Methods Research*.

El desarrollo ha sido tan rápido que la *ISI Web of Knowledge* recoge 2800 trabajos publicados que incluyen el término "*Mixed Method*" como tema principal. Estos trabajos se enmarcan principalmente en tres áreas: "investigación educativa", "salud pública" y "psicología". Centrándonos en el área de "psicología" se observa que los 294 estudios situados en este área fueron publicados a partir de 1998, punto en el que puede considerarse, también a nivel teórico, el inicio de la época más productiva. Además, se observa un incremento de los trabajos en los últimos años, ya que 157 estudios (más del 50%) se han publicado entre 2010 y 2012. En cuanto a la aparición de otros términos relevantes en estos trabajos, 18 estudios recogieron dentro de las palabras clave el término "*validity*" y ocho la expresión "*test development*". Además, el número de investigadores implicados en el desarrollo de la investigación mixta aumenta como demuestra la *Mixed Method International Conference* que reúne anualmente a los profesionales del área y celebrará en 2012 su octava edición (<http://www.healthcareconferences.leeds.ac.uk>).

Los datos bibliométricos muestran la importancia de un movimiento en el cuál se recogen múltiples perspectivas, porque aunque todo estudio clasificado como investigación mixta implementa varios procedimientos, la forma de hacerlo es muy versátil. Versatilidad que se refleja en la variedad de diseños posibles que han sido clasificados siguiendo diferentes criterios. Una de las clasificaciones de los diseños más utilizada es la realizada por Creswell (1995) que recoge dos dimensiones: la secuencialidad y la dominancia. Según la

secuencialidad, los estudios pueden ser simultáneos o secuenciales, es decir, pueden aplicar los dos tipos de métodos paralelamente o en fases sucesivas; mientras que la dominancia describe la prioridad, pudiendo ser más dominante la parte cuantitativa, la parte cualitativa o ambas igualmente relevantes. Combinando estas dos dimensiones se obtienen seis tipos de diseños básicos, que se convierten en nueve si se considera el orden, factor que afectaría solamente a los estudios secuenciales.

Creswell (1995) también describe un sistema de representación combinando tipos de letras y signos de puntuación. Por ejemplo, "QUAN + qual" indica que el estudio se realiza de manera secuencial siendo la parte cuantitativa la primera en ejecutarse y la más dominante. "QUAN/QUAL" indica que ambas partes fueron simultáneas y ninguna de ellas fue dominante. Considerando las características del estudio, el diseño anterior (QUAN/QUAL) podría ser, por ejemplo, un estudio que aplica un análisis cuantitativo para analizar las preguntas de elección múltiple de una encuesta y un análisis cualitativo para las preguntas abiertas de esa misma encuesta. Mucho más frecuentes son las investigaciones que incluyen dos mini-estudios, uno aplicando una perspectiva cuantitativa sobre unos datos y otro aplicando una perspectiva cualitativa sobre otros datos diferentes, que son posteriormente combinados con el fin de extraer conclusiones conjuntas.

Las situaciones descritas son ejemplos de lo que específicamente se denomina Métodos Mixtos o *Mixed Methods*. Sin embargo, dentro de la investigación mixta existen otras variantes, como los denominados Estudios de Modelos Mixtos o *Mixed Model Studies* (Tashakkori y Teddlie, 1998). Esta categoría incluye estudios que aplican metodologías cuantitativas y cualitativas en la misma fase, por ejemplo estudios que analizan los datos cualitativamente elaborando categorías que luego se resumen en forma de frecuencias o tablas de contingencia. El último grupo de estudios recogidos bajo la denominación de investigación mixta son aquellos que, como indica Brannen (2005), implementan dos procedimientos aunque ambos formen parte de un mismo

paradigma. Tashakkori y Teddlie (1998) los denominan estudios de Método Único o *Monomethod*. En las tres variantes, aunque el diseño cambie, el objetivo siempre es acceder a la mayor cantidad de información y que ésta información sea de la mayor calidad posible.

En el contexto de la investigación mixta se han planteado necesidades relacionadas con situaciones en que los datos obtenidos con un tipo de método, por ejemplo cualitativo, se utilizan para completar datos procedentes de otro método, por ejemplo cuantitativas. Dentro del paradigma cualitativo, los métodos de pretest cognitivo son unos de los métodos más utilizados en los últimos años. Su flexibilidad y facilidad de uso han provocado que su aplicación sea cada vez más amplia y variada. Entre los métodos de pretest cognitivo, las Entrevistas Cognitivas (EC) son uno de los métodos más conocidos e implementados actualmente, lo que lleva a plantearse su utilización en contextos diferentes y a preguntarse, ¿podrían los métodos de pretest cognitivo aplicarse para resolver alguno de los problemas emergentes en el campo aplicado?, por ejemplo, ¿podrían utilizarse métodos de pretest cognitivo para interpretar resultados cuantitativos?, o ¿podrían servir las Entrevistas Cognitivas (EC) para ayudarnos a explicar los “números” que nos proporcionan los análisis psicométricos?

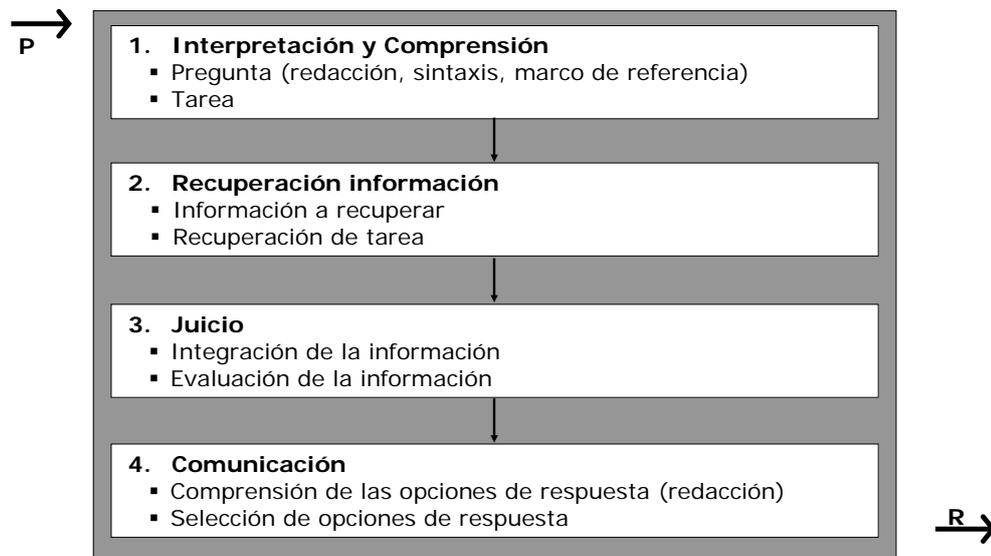
A continuación, se presentará una breve introducción a los métodos de pretest cognitivo con el fin de facilitar la comprensión del enfoque metodológico que se utiliza para diseñar los estudios incluidos en esta tesis. Se mostrará como los métodos de pretest cognitivo pueden facilitar la realización de estudios de validación dentro del paradigma de investigación mixta.

Los métodos de pretest cognitivo

Los métodos de pretest cognitivo nacen de la combinación de dos factores: la insatisfacción de los investigadores de encuesta por el tratamiento que se daba rutinariamente a los errores de medición, clasificados como “errores ajenos al muestreo” por Groves (1989), y el cambio de paradigma que supuso la irrupción de la psicología cognitiva y el interaccionismo simbólico (Foddy, 1996). Ambos factores incitaron el cambio desde el modelo automático, casi de estímulo-respuesta, hasta un modelo que contemplaba el comportamiento de la persona que responde al cuestionario, es decir, un modelo más complejo para explicar el “proceso cognitivo” implicado en el proceso de respuesta (Willis, 2005).

Sin duda, el modelo más citado para ilustrar la nueva concepción sobre la presencia de los procesos cognitivos que ocurren entre la pregunta y la respuesta, es el modelo cognitivo “pregunta-y-respuesta” que caracteriza el denominado movimiento *Cognitive Aspects of Survey Methodology* (CASM). El movimiento surgió como resultado de dos conferencias: el *Advanced Research Seminar on Cognitive Aspects of Survey Methodology* celebrada en 1983 en Estados Unidos; y la *Conference on Social Information Processing and Survey Methodology* que tuvo lugar en 1984 en Alemania (Jabine, Straf, Tanur y Tourangeau, 1984). En ambas conferencias se destacó la importancia de tener en cuenta a las personas como “agentes activos”, y contemplar los procesos cognitivos que intervienen en ese proceso de respuesta. Tourangeau (1984) desarrolló posteriormente el modelo del proceso “pregunta-y-respuesta” que incluía, además de los elementos tradicionales, una descripción de los procesos cognitivos implicados en el proceso “pregunta-y-respuesta”. La Figura 1 muestra una representación del modelo en la que se sitúan los procesos cognitivos que aparecen entre la formulación de la pregunta y la emisión de la respuesta.

Figura 1. Representación modelo pregunta-y-respuesta



La figura 1 muestra las cuatro fases que una persona completaría cuando responde a una pregunta. Durante estas cuatro fases que ocurren entre la formulación de la pregunta y la emisión de la respuesta, el participante desarrolla varias operaciones cognitivas: primero interpreta y comprende la pregunta o tarea planteada lo que implica comprender tanto el objetivo pretendido como los conceptos y expresiones que incluye; a continuación, recupera la información necesaria para responder a la pregunta; después realiza un juicio que le permite integrar y evaluar la información recuperada; y finalmente, ajusta su respuesta a las alternativas propuestas y la comunica.

En las dos últimas décadas, el modelo ha ido evolucionando para incluir la posible no secuencialidad de las fases en todas las circunstancias (Collins, 2003), y, la presencia de dimensiones sociales y culturales. Las dimensiones sociales y culturales son claves para explicar la motivación tanto del entrevistador como del entrevistado en el proceso de “pregunta-y-respuesta”. Por ejemplo, una misma pregunta puede llevar a diferentes respuestas según se formule en un contexto formal o informal. Krosnick (1999), señaló la importancia de los aspectos motivacionales destacando la posible influencia de los llamados procesos de “optimización” y “satisfacción”. Estos procesos hacen referencia a

las motivaciones, ya sean intrínsecas o extrínsecas, que cada participante tiene para responder a una pregunta. La optimización requiere realizar todas las fases del proceso y hacerlo de forma completa. La alternativa a la optimización es la satisfacción. Un encuestado “satisface” cuando salta alguna fase del proceso “pregunta-y-respuesta” o las completa de forma superficial. La satisfacción aparece en aquellos participantes que quieren que la interacción entrevistador-entrevistado termine de forma “satisfactoria” para ambos. Factores como la dificultad de la pregunta, el nivel educativo de los entrevistados, la existencia o no de recompensas, etc., han mostrado tener una influencia consistente en los niveles de “satisfacción”.

El objetivo de los procedimientos de pretest cognitivo es indagar en el proceso “pregunta-y-respuesta” para obtener evidencias con una doble finalidad: “optimizar” las preguntas e inferir lo que “realmente” están midiendo.

Algunos de los métodos de pretest cognitivo más utilizados son las Entrevistas Cognitivas (EC), la Codificación del Comportamiento (CC), los debriefing o la expert appraisal (Presser et al. 2004). Nos centraremos en la descripción de los dos primeros por ser los aplicados en los estudios que se incluyen en esta tesis.

Entrevistas cognitivas

Las Entrevistas Cognitivas (EC) son el método de pretest cognitivo más utilizado en los estudios que buscan conocer el proceso de “pregunta-y-respuesta” realizado por los participantes (Willis, 2005). El éxito de este método se debe en parte a su flexibilidad y a la capacidad para adaptarse a los objetivos del investigador. La utilización de las EC se ha realizado mayoritariamente en los institutos oficiales de estadística, donde se ha seguido una aproximación propia de la psicología cognitiva. Sin embargo, es necesario mencionar otro enfoque más sociológico de las EC, en el que se considera la presencia

mediadora de factores sociales y culturales, dando al entrevistado un rol más centrado en la traslación de sus experiencias vitales al proceso de “pregunta-y-respuesta” (Miller, Chepp, Willson y Padilla, 2012, en prensa).

Entre las numerosas definiciones que surgen, desde la perspectiva más cognitiva, de lo que es una Entrevista Cognitiva, una de las más completas y consensuadas es la enunciada por Beatty y Willis (2007), que la describen como “la administración de un borrador de las preguntas de la encuesta que recoge información verbal adicional sobre las respuestas a las preguntas de la encuesta, dicha información es usada para evaluar la calidad de las respuestas o para ayudar a determinar si la pregunta está generando la información que el autor de la encuesta pretendía” (p. 288). La definición aboga por una determinación previa de las intenciones del investigador, que se traslada a un protocolo de entrevista para comprobar el grado de ajuste entre lo pretendido y lo alcanzado.

El protocolo de entrevista puede desarrollarse siguiendo un método *think-aloud* (Ericson y Simon, 1980) centrado en la verbalización de los pensamientos de los participantes mientras responden a las preguntas objetivo, o siguiendo el método *probing based* (Willis, DeMaio y Harris-Kojetin, 1999), que elabora pruebas de indagación para aspectos concretos de cada pregunta. En este último caso, las pruebas se desarrollan en función de las características o elementos de las preguntas que los investigadores consideren potencialmente problemáticos. Por ejemplo, cuando se sospecha que el significado de una palabra puede ser confuso, se puede desarrollar una prueba de indagación para conocer la interpretación realizada por los entrevistados. La información obtenida sobre las diferentes interpretaciones posibles permitirá realizar ajustes que faciliten la transmisión del concepto previsto. La Tabla 1 muestra seis tipos de pruebas de indagación especificados por Willis (2005), así como un ejemplo para ilustrar el objetivo de cada una.

Tabla 1. Tipos y ejemplos de pruebas cognitivas.

Prueba Cognitiva	Ejemplo
Prueba General	¿En qué estaba pensando mientras respondía a la pregunta? ¿Qué información le ha venido a la cabeza mientras respondía?
Prueba sobre Comprensión de conceptos	¿Qué ha entendido por el término “salud”?
Prueba de Parafraseo	¿Podría repetir la pregunta con sus propias palabras?
Juicios de Confianza	¿Con qué seguridad afirma que... “ha acudido al médico más de cinco veces en el último año”?
Prueba sobre Recuperación de Información	¿En qué se ha basado para decir que... “acudió al médico más de cinco veces durante el año pasado”?

También existen diferentes aproximaciones al análisis de los datos procedentes de las EC. Por un lado, aparecen modelos basados en esquemas de codificación. Estos modelos resumen los resultados de las EC utilizando generalmente entre dos y nueve categorías generales (Willis, DeMaio, y Harris-Kojetin, 1999). Las categorías se desarrollan en función de las características y objetivos de la investigación, aunque una tendencia habitual es plantear una agrupación que responda a lo que el investigador “estaba buscando”, que es posteriormente completada con lo que el investigador “ha encontrado” sin planteárselo. Por ejemplo, Willis (2005) propone cinco categorías generales para guiar una primera revisión de los resultados. Estas categorías son:

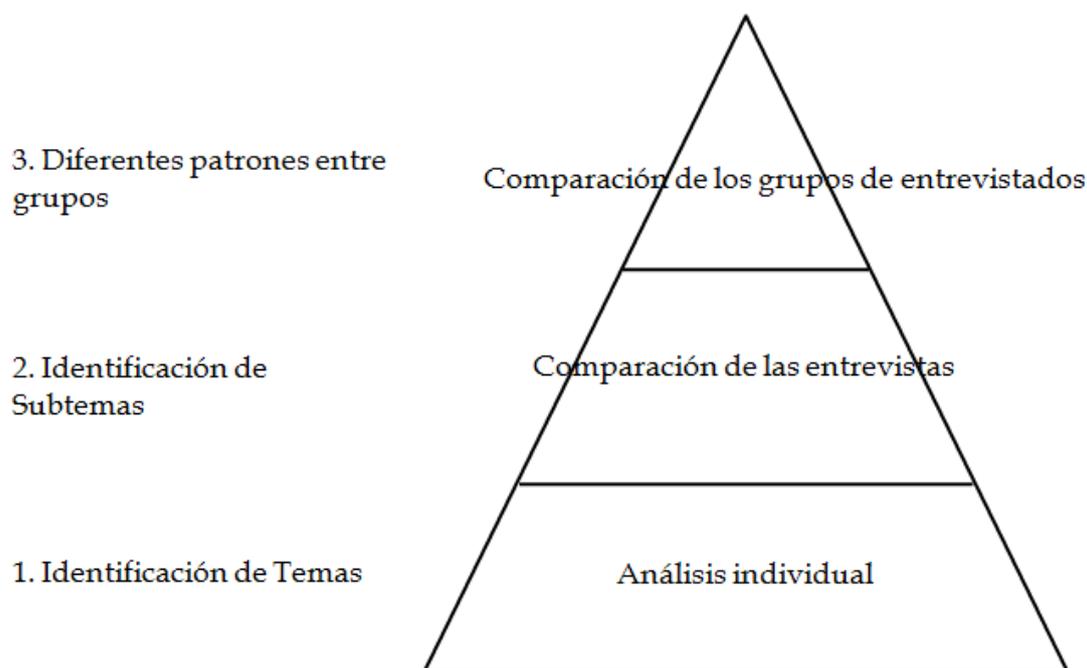
- a) Recomendaciones específicas del ítem, dirigidas a los términos y estructura utilizados;
- b) Especificación de los objetivos, en relación a la capacidad del ítem para satisfacerlos;

- c) Problemas relacionados con el orden y otras interacciones entre los ítems;
- d) Problemas relacionados con la longitud del instrumento, y;
- e) Limitaciones sobre lo “preguntable”, es decir, a qué informaciones podemos acceder con los métodos disponibles.

Dentro de los modelos basados en esquemas de codificación, puede plantearse un análisis más detallado especificando las categorías descritas. Para ello, se realizaría en primer lugar una revisión de las narraciones surgidas ante el protocolo. A continuación, se determinarían posibles categorías objetivas basadas en los elementos clave de esas narraciones y se asignarían dichas categorías a los segmentos de las narraciones. Por ejemplo, podrían clasificarse los segmentos procedentes del protocolo agrupando lo ocurrido durante la entrevista en función de la fase a la que pertenezca dentro del modelo “pregunta-y-respuesta”. Es decir, podrían dividirse los segmentos de las narraciones según muestren la interpretación que el participante ha hecho de los conceptos incluidos en la pregunta o el proceso realizado para ajustar su respuesta a las alternativas propuestas.

Desde la aproximación sociológica, se han desarrollado estrategias de análisis más cualitativas centradas en el discurso y narrativas de los entrevistados. Por ejemplo, Miller (2007) desarrolló un modelo por etapas que recoge tres fases en el análisis de las EC. La Figura 2 muestra la representación gráfica más reciente de este modelo que incorpora algunas modificaciones sobre la propuesta original (Miller, 2007).

Figura 2. Representación gráfica del modelo piramidal



Como muestra la Figura 2, las tres fases del análisis van desde los aspectos más elementales y ligados a los datos, hasta los aspectos más complejos situados en la cúspide. En primer lugar, el análisis individual de las entrevistas permite obtener los denominados temas (o *themes*). Los temas se refieren a las temáticas generales que desarrollan los participantes. Por ejemplo, si a una persona le preguntamos qué ha pensado para responder a una pregunta sobre actividades cotidianas, esa persona puede enumerarnos lugares, situaciones, personas, etc. Estas categorías generales serían los temas. El objetivo del segundo nivel es comparar las entrevistas para obtener sub-temas (o *sub-themes*). Los sub-temas son especificaciones de los temas generales. Por ejemplo, en el caso anterior podríamos dividir el tema “personas” en “familiares” y “no familiares” o en “familiares”, “amigos” y “compañeros de trabajo” dependiendo de cuál sea nuestro objetivo. Por último, en el tercer nivel se busca la comparación entre distintos grupos con el fin de detectar patrones de interpretación diferentes y relacionados con determinadas características de los participantes. En el ejemplo anterior, un resultado propio del tercer nivel podría

ser encontrar que las mujeres usan más frecuentemente el sub-tema “familiares” y los hombres el sub-tema “compañeros de trabajo”.

La aportación de este modelo es la diferenciación de los hallazgos en estos tres niveles que permiten plantear distintos tipos de conclusiones, es decir, que los resultados extraídos en cada nivel de la pirámide pueden emplearse para alcanzar distintos objetivos. Los resultados de la fase 1 pueden por ejemplo guiar la modificación o corrección de algunas preguntas de forma que se asegure el ajuste de la interpretación a las intenciones del investigador, mientras que los resultados de la fase 3 pueden utilizarse para establecer diferencias en la interpretación de constructos a través de distintos grupos poblacionales. Pero, ¿podrían utilizarse estos resultados para interpretar diferencias entre grupos detectadas mediante otros procedimientos como puede ser un análisis del Funcionamiento Diferencial de los Ítems (DIF)? ¿Podrían usarse para enriquecer y complementar los resultados obtenidos mediante procedimientos psicométricos? Estas preguntas son algunas de las que se han planteado en los estudios de esta tesis.

Codificación del comportamiento

La Codificación del Comportamiento (CC) desarrollada por Cannell en los 60 (Cannell, Fowler, y Marquis, 1968), se centra en otro aspecto considerado básico para localizar características o elementos problemáticos de las preguntas: la interacción entrevistador-entrevistado. La CC es un procedimiento basado en la observación sistemática de las entrevistas. Consiste en la codificación de los hechos ocurridos durante la interacción que tiene lugar entre el entrevistador y el entrevistado cuando se administra un cuestionario. Para realizar la codificación pueden incluirse diferentes categorías de observación en función de los intereses de los investigadores. Además, existen esquemas de codificación que resumen los aspectos claves que deben ser registrados sobre el entrevistador y el entrevistado. Uno de ellos es el esquema de codificación de

Oksenberg, Cannell y Kalton (1991), que describe distintas categorías definidas por el actor (entrevistador o entrevistado), y el momento (durante la lectura de la pregunta o durante la emisión de la respuesta). La combinación de las categorías da lugar a distintos códigos que recogen los comportamientos más relevantes ocurridos durante la interacción entrevistador-entrevistado. La Tabla 2, muestra algunos de los comportamientos incluidos en las clasificaciones habituales, así como una breve descripción.

Tabla 2. Categorías para la clasificación de los comportamientos ocurridos durante la interacción entrevistador-entrevistado

Actor	Momento	Códigos	Descripción
Entrevistador	Durante la presentación de la pregunta	Lectura exacta	El entrevistador lee la pregunta tal y como está escrita en el guión
		Ligeros cambios	El entrevistador lee la pregunta cambiando algún detalle pero sin alterar el significado de la pregunta
		Grandes cambios	El entrevistador introduce cambios que alteran el significado de la pregunta o no completa su lectura
Entrevistado	Durante la presentación de la pregunta	Interrupción	El entrevistado interrumpe la lectura para preguntar algo o para responder a la pregunta
		Solicitud de repetición	El entrevistado pide que se le repita la pregunta
		Solicitud de clarificación	El entrevistado pide que se le aclare algún aspecto de la pregunta
	Durante la emisión de la respuesta	Cambia la respuesta	El entrevistado modifica la respuesta tras una clarificación o por otro motivo
		Expresa dudas o malestar	El entrevistado duda al responder o se siente incómodo ante la pregunta

Actor	Momento	Códigos	Descripción
		Respuesta no ajustada	La respuesta no coincide con ninguna de las alternativas ofrecidas aunque su contenido refleja la comprensión pretendida de la pregunta
		Respuesta inválida	La respuesta no está relacionada con la pregunta
		Respuesta no sé	El entrevistado no sabe la respuesta
		Respuesta limitada	La respuesta indica incertidumbre o es poco precisa
		Respuesta adecuada	La respuesta se ajusta al objetivo de la pregunta

Los comportamientos recogidos en la Tabla 2 se dividen en primer lugar en función de la persona que las realiza: entrevistador o entrevistado. Las conductas más relevantes del entrevistador ocurren durante la lectura de la pregunta, ya que es el momento en que el entrevistador puede realizar modificaciones sobre el guión establecido. Sin embargo, las conductas relevantes por parte del entrevistado pueden ocurrir tanto durante la presentación de la pregunta como durante la emisión de la respuesta. Durante esta última fase suele tener lugar la respuesta del participante que también es codificada en función del grado en que cubre los objetivos previstos.

Dependiendo de las conductas ocurridas durante la presentación de la pregunta y durante la emisión de la respuesta, las secuencias pueden ser clasificadas. Una secuencia es todo lo que ocurre desde la presentación de una pregunta hasta la presentación de la pregunta siguiente. Siguiendo las indicaciones de Ongena y Dijkstra (2006) existen tres tipos de secuencias. Las secuencias paradigmáticas son aquellas consideradas “ideales”, es decir, en las que la pregunta es leída por el entrevistador tal y como se recoge en el guión, la respuesta del entrevistado es adecuada y el entrevistador la reconoce como tal (lectura exacta + respuesta adecuada). Las desviaciones de esta secuencia “ideal” son denominadas secuencias no paradigmáticas, y éstas, a su vez,

pueden ser clasificadas como: problemáticas o no problemáticas. En el primer grupo se engloban aquellas que tienen efectos negativos sobre los datos; por ejemplo, que el entrevistador haga una lectura con grandes cambios o que el entrevistado proporcione una respuesta no ajustada. Entre las no problemáticas se encuentran aquellas secuencias en las que ha ocurrido algo problemático pero ha sido resuelto o no ha influido negativamente en los datos. Un ejemplo de esta situación sería cuando el participante solicita que se le repita la pregunta porque ha ocurrido algún ruido inesperado que le ha impedido escuchar el enunciado.

Tradicionalmente, los resultados obtenidos mediante la CC se han empleado para localizar posibles aspectos problemáticos durante la interacción y así señalar los aspectos a los que se deben prestar atención. Por ejemplo, los resultados de la CC pueden mostrar la necesidad de incluir la definición de algún concepto cuando en una pregunta los participantes han solicitado frecuentemente una aclaración. Pero, ¿podría utilizarse la CC en otros contextos? Por ejemplo, ¿podría emplearse la CC, con un carácter más interpretativo, para evaluar la calidad de las respuestas de informantes con unas características concretas? ¿Podría también proporcionar otros datos como la convergencia entre las respuestas de distintos tipos de informantes aportando una depuración más precisa de los datos que la ofrecida por otros métodos? Algunos de los estudios incluidos en esta tesis mostrarán datos relevantes para resolver estos planteamientos.

Una vez situado el marco metodológico desde lo más general hasta lo más específico, trataremos los problemas de investigación objeto de los estudios de esta tesis, con el fin de ilustrar los motivos que han llevado a desarrollar cada uno de ellos.

Los informantes indirectos o proxies.

Comenzaremos por situar la problemática de los informantes indirectos o proxies en el contexto de la investigación mediante encuesta, donde habitualmente el cuestionario que se pretende aplicar va dirigido a una población específica, es decir, se diseña pensando en las personas que lo responderán. El problema es que en algunas ocasiones no es posible localizar en el domicilio al informante directo o persona “objetivo” de la encuesta o, dicha persona, no puede responder por su condición de salud o restricciones legales. Esta situación plantea dificultades en términos de organización y de costes, lo que provoca la necesidad de buscar soluciones. Una posible solución es pedir a otra persona (proxy), que sí esté presente en el domicilio, que responda las preguntas poniéndose en el lugar de la persona a la que se pretendía administrar el cuestionario (informante directo). De esta manera se evita el hecho de tener que volver en otro momento, sustituir a la persona objetivo, o perder la información de un participante aumentando la tasa de no-respuesta.

La consecuencia obvia de tomar la decisión de entrevistar a un proxy, es que la información que se obtiene no es la información directa que se buscaba, por lo que se deben contemplar posibles discrepancias entre las distintas fuentes de información. La utilización de “proxies” ha recibido muchas críticas por parte de los investigadores, lo que ha llevado a la *Eurostat Task Force* a limitar el uso de los proxies a situaciones estrictamente necesarias como son que los informantes directos no puedan responder por motivos de salud o por motivos legales (Eurostat Task Force, 2005; Tafforeau, Lopez, Tolonen, Scheidt-Nave, y Tinto, 2006).

Sin embargo, las situaciones en que no se tiene acceso al informante directo siguen estando presentes y la utilidad de los proxies sigue siendo clara en otros escenarios cotidianos fuera de esas limitaciones legales o personales. Uno de esos escenarios puede proceder de la frecuente incompatibilidad horaria entrevistador-entrevistado por la longitud o las características de la jornada

laboral de uno de ellos o de ambos. El hecho de “necesitar” la utilización de los proxies provoca que los investigadores responsables de la administración de encuestas se planteen indagar la calidad real de las respuestas proporcionadas por este tipo de informantes, así como la convergencia entre las respuestas de estos y las de los informantes directos. La evaluación de la convergencia entre las respuestas de los informantes directos y los proxies, y la evaluación de la calidad de las respuestas de los proxies han sido tradicionalmente estudiadas de forma paralela. La calidad de las respuestas de los proxies se considera adecuada cuando coincide con la emitida por el informante directo, es decir, la respuesta del informante directo es considerada como el “gold standard”.

Al analizar la información obtenida en los estudios que utilizan proxies, se observa que en la mayoría se evalúa la convergencia de las respuestas en base exclusivamente a índices estadísticos y olvidando los aspectos cualitativos (Pickard et al., 2004; Todorov y Kirchner, 2000). Cuando esta evaluación es realizada exclusivamente a partir de resultados estadísticos, se puede provocar la pérdida de información relevante relacionada con el contenido de las respuestas, además de una comparación entre datos incomparables. Por ejemplo, si usamos sólo índices estadísticos necesitaremos transformar las respuestas y codificarlas para poder compararlas “estadísticamente”. Esta codificación se realiza habitualmente asignando números a las categorías de respuesta. Cuando se sigue este procedimiento, situaciones como la *no respuesta* o el *no sé* se codifican como respuesta nula, por ejemplo, con un cero. Pero, ¿qué pasa cuando comparamos una respuesta nula con cualquier otra respuesta? En este caso estamos ante una situación de desacuerdo que sería numéricamente representada de la misma manera que otra situación en la que dos personas dieran respuestas válidas pero diferentes. Es decir, una persona responde “sí” y otra “no” nos encontramos ante una situación de desacuerdo que, hasta ahora, se ha hecho equivalente a otra situación en la que una persona responde “sí” y la otra dice “no sé”. Esta amenaza a la comparabilidad se debe a que lo frecuente ha sido determinar o etiquetar lo ocurrido como “acuerdo” o “desacuerdo” entre informantes, más que analizar la naturaleza del desacuerdo.

Sin embargo, en algunos contextos puede ser relevante la diferenciación entre los tipos de desacuerdos. De hecho, la diferenciación se plantea como imprescindible cuando queremos localizar preguntas a las que un grupo de informantes no han respondido sistemáticamente, mientras que sí se ha obtenido respuesta de los participantes del otro grupo. En la otra vertiente, puede ser importante distinguir diferencias entre respuestas válidas cuando estemos interesados en detectar contradicciones entre informantes en una pregunta por ejemplo de "Si / No". Por lo tanto, necesitamos algo más que esa etiqueta de "desacuerdo". Esta información, que puede resultar valiosa, no la obtenemos con los métodos que actualmente se aplican para evaluar la convergencia y, sin embargo, podría ser accesible si se añaden, a la codificación habitual de las respuestas, categorías sustantivas que nos den más información sobre lo que está ocurriendo.

Otro aspecto problemático en la investigación con proxies es la selección de éstos. Cuando un entrevistador llega a un domicilio buscando a su informante directo y no lo encuentra, lo más rentable y cómodo es utilizar como proxy a cualquiera de las personas presentes en ese momento. Sin embargo, además de las variables sociodemográficas controladas habitualmente, aspectos que no suelen tenerse en cuenta, como el tipo de relación entre el proxy y el informante directo, pueden afectar a la calidad de los datos; es decir, no es lo mismo que el proxy sea un hermano del informante directo que sea su compañero de piso. También son relevantes aspectos como el contenido de la pregunta. Por ejemplo, se ha observado que los proxies dan respuestas más adecuadas cuando se les pregunta acerca de áreas de salud observables y compartidas con otras personas, tales como, problemas con actividades físicas de la vida cotidiana o limitaciones crónicas (Magaziner, Bassett, Hebel, y Gruber-Maldini, 1996); mientras que sus respuestas son más limitadas cuando se les pregunta por el estado emocional, el dolor y otras dimensiones de salud no directamente observables (Grootendorst, Feeny, y Furlong, 1997).

En resumen, se plantean varias situaciones o necesidades que podrían cubrirse si conseguimos mejorar el proceso para evaluar la calidad de las respuestas de los proxies y el análisis de la convergencia entre las respuestas de distintos informantes. Esto, además, podría contribuir a aportar claridad sobre las características que deben tener las preguntas diseñadas para ser respondidas por los proxies o incluso sobre las características del “proxy ideal”; elementos que, como indican Rajmil, et al. (1999) son, junto la relación existente entre los informantes, los principales factores de influencia en las respuestas de los proxies.

La mejora del proceso implica, por un lado, evitar evaluaciones basadas exclusivamente en resultados cuantitativos, completando estos datos con información cualitativa que pueda resultar relevante y, por otro, controlar comparaciones no equivalentes entre las respuestas de los participantes. Llegamos así a las cuestiones que han sido abordadas en los estudios de la tesis: ¿es posible evaluar la calidad de las respuestas de los proxies por sí mismas, es decir, dejando a un lado la respuesta de los informantes directos aunque esto sea el estándar de comparación?, es decir, ¿podemos saber si las respuestas de los proxies son buenas aunque no conozcamos las respuestas de los informantes directos? ¿Tiene algún efecto, en la comparación del acuerdo, el tipo de relación que tienen los informantes? Por otra parte, ¿podría esta evaluación incluir información cualitativa que nos ayudara a conocer el origen del desacuerdo entre los informantes? ¿Podría purificarse el análisis del comportamiento de los participantes de forma que conociéramos el tipo de “desacuerdo” entre informantes?

Los análisis psicométricos y el Funcionamiento Diferencial de los Ítem (DIF).

Llegamos ahora al último punto que debemos abordar para poder enmarcar todos los trabajos incluidos en esta tesis. En este último apartado se introducirán algunos aspectos sobre los análisis psicométricos y el Funcionamiento Diferencial de los Ítems (DIF) que facilitarán la comprensión del objeto de estudio planteado en los últimos trabajos de la tesis.

Estamos tan familiarizados con los análisis psicométricos que, en ocasiones, los aplicamos de forma rutinaria para obtener información sobre la calidad de los instrumentos que administramos. Algunos de los índices estadísticos más utilizados en todas las disciplinas son la media o la varianza, que nos describen la distribución de las respuestas y su variabilidad; o el índice de discriminación (correlación ítem-total corregida), que nos cuantifica la capacidad que tiene un ítem para diferenciar entre los niveles de la variable que muestran las personas. También el coeficiente alfa o el análisis factorial son procedimientos ampliamente utilizados para estimar la fiabilidad y describir la estructura interna del instrumento. Estos análisis permiten conocer las debilidades y fortalezas generales del cuestionario o test que estamos utilizando. Sin embargo, a veces obtenemos resultados inesperados que no sabemos interpretar o cuyas causas no podemos explicar. Esto mismo ocurre en el análisis del DIF, aunque en este caso las implicaciones son mayores porque puede afectar a las conclusiones obtenidas, minando la validez de las interpretaciones y/o perjudicando a un grupo sobre otro.

El DIF se produce cuando participantes con idéntico nivel en la característica medida, es decir, comparables, tienen distintas probabilidades de respuesta para un determinado ítem dependiendo del grupo al que pertenezcan (Millsap y Everson, 1993). Cuando analizamos el DIF comparamos dos grupos que tradicionalmente han sido denominados como Grupo de Referencia (GR) y Grupo Focal (GF). La designación de los grupos puede realizarse utilizando criterios tradicionales, como son considerar al grupo más numeroso o a aquel

que responde a la versión original del instrumento como el GR; o respondiendo a los objetivos del estudio concreto, como por ejemplo utilizando como GR el grupo objeto de interés. Según la distribución del DIF, éste puede ser uniforme o no uniforme entre los grupos, dependiendo de que exista interacción entre el atributo medido y la pertenencia a un grupo (Mellenbergh, 1982). Existe una gran variedad de técnicas estadísticas para evaluar el DIF cuyo funcionamiento ha sido ampliamente estudiado en los últimos años (Zumbo, 2007; Hidalgo y Gómez-Benito, 2010). Entre ellas, el estadístico Mantel-Haenzsel (MH) y la Regresión Logística (RL) son los procedimientos más utilizados y los que se aplicaron en el Estudio 4 de esta tesis donde son descritos más detalladamente.

En apartados anteriores se discutió sobre la importancia de que las metodologías existentes sean capaces de adaptarse y dar respuesta a los problemas de investigación que van surgiendo en cada momento histórico. El DIF es un buen ejemplo de esto, ya que su auge se debe en parte al esfuerzo por responder a las necesidades que se plantean en el contexto de los estudios transculturales. El hecho de que cada vez aparezcan más situaciones en las que se necesita comparar grupos, hace que análisis como el DIF ganen importancia y su desarrollo sea abordado con más profundidad. El DIF permite señalar aquellos aspectos problemáticos, que por estar presentes en el instrumento que estamos usando, pueden perjudicar a un grupo haciéndonos llegar a conclusiones no reales. La detección del DIF permite localizar características del ítem que, por estar funcionando de forma diferente en los grupos implicados, pueden provocar diferencias no relacionadas con las diferencias reales en el atributo medido. Por ello, el DIF nos advierte sobre la precaución que debemos tomar al establecer conclusiones y sobre los problemas que debemos resolver para poder tener seguridad en la equivalencia entre los grupos.

Por la preocupación que ha despertado esta amenaza a la validez de las comparaciones entre grupos, el número de estudios que aplican análisis del DIF para evaluar instrumentos ha crecido notablemente en los últimos años (Cho, Martin, Conger, y Widaman, 2010; Kalaycioglu, y Berberoglu, 2011; Lewis,

Yang, Jacobs, y Fitchett, 2012). Sin embargo, la evaluación de qué es lo que provoca el DIF no ha tenido resultados tan exitosos. A pesar de que la investigación de las causas del DIF se ha intentado abordar desde distintos frentes, por ejemplo a través del estudio de las características de los instrumentos o de los ítems, o a partir de la evaluación de la calidad de la adaptación, aún no se han alcanzado resultados concluyentes. Por ello, en los últimos tiempos, se ha intentado abordar el estudio de las causas del DIF desde perspectivas más novedosas como mediante la aplicación de técnicas cualitativas o la utilización de modelos multinivel (Swanson, Clauser, Case, Nungester, y Featherman, 2002; Van den Noortgate, De Boeck, y Meulders, 2003).

Aunque los esfuerzos han sido cada vez más intensos, aún no se ha conseguido llegar al descubrimiento de las causas del DIF y a su origen. Nuestra propuesta radica en la aplicación de métodos de pretest cognitivos para contribuir en la indagación de esas causas. Se plantea evaluar la utilidad de estos métodos para ayudarnos a interpretar estos resultados inesperados que nos ofrecen los análisis psicométricos y que no podemos explicar. Además, se plantea la alternativa de utilizar las EC para explicar e interpretar los resultados del DIF y para esclarecer la búsqueda de sus causas.

Una vez desarrolladas de una forma muy básica las temáticas tanto metodológicas como conceptuales tratadas en esta tesis, pasaremos a exponer los objetivos y la estructura de los estudios incluidos.

OBJETIVOS Y ESTRUCTURA

Concretando lo visto en la introducción, queda claro que uno de los pilares de la presente tesis es la convergencia entre los métodos cualitativos y cuantitativos para evaluar la calidad de las mediciones en diferentes contextos. El objetivo que ha marcado el proceso durante el cual se ha realizado este proyecto, ha sido afrontar problemas metodológicos asociados a la calidad de la información que aportan cuestionarios y escalas, para obtener evidencias de validez mediante la aplicación de métodos de pretest cognitivos y métodos cuantitativos, especialmente, psicométricos. Para alcanzar este objetivo general, se plantean los siguientes objetivos específicos que se abordarán a partir de la realización de estudios concretos:

- Objetivo específico 1: Conocer la capacidad de la Codificación del Comportamiento (CC) para evaluar las respuestas de los proxies. Estudios 1 y 2.
- Objetivo específico 2: Evaluar la utilidad de las Entrevistas Cognitivas (EC) como procedimiento para interpretar resultados psicométricos. Estudios 3 y 5.
- Objetivo específico 3: Determinar la capacidad de las EC para localizar las causas del DIF. Estudios 4 y 5.

A su vez, los estudios desarrollados para alcanzar estos objetivos se pueden describir utilizando tres dimensiones: el método cognitivo aplicado, el diseño de la investigación y la forma de combinar los resultados. Situemos cada uno de los estudios teniendo en cuenta esta información.

En primer lugar, en cuanto al método cognitivo empleado, en dos de los estudios se aplicó la CC (Estudio 1 y Estudio 2) y en dos de ellos las EC (Estudio 3 y Estudio 5). En los dos primeros estudios, la CC permitió evaluar la precisión de las respuestas proporcionadas por los proxies a partir del análisis de la interacción entrevistador-entrevistado. En el Estudio 1 se evaluó la calidad de las respuestas proporcionadas por los proxies, mientras que en el Estudio 2 se observó la convergencia entre las respuestas de informantes proxies e informantes directos. Por otra parte, en los Estudios 3 y 5 se aplicaron las con el

fin de obtener información sobre las interpretaciones realizadas por los participantes al responder a los ítems de una escala. En el Estudio 4 no se aplicó directamente ninguno de los métodos de pretest cognitivo, sin embargo, el diseño de este estudio se realizó con el objetivo de obtener información que permitiera elaborar las pruebas de indagación del protocolo de entrevista cognitiva utilizado en el Estudio 5. A su vez, la información proporcionada por las EC permitió interpretar los datos de este estudio en el que se analizó el DIF, por lo que podría situarse dentro del segundo grupo de estudios.

La segunda dimensión se refiere al diseño de la investigación. Tras la tradicional división de los paradigmas en cuantitativos y cualitativos, surge la investigación mixta, y dentro de ella se describen tres tipos de estudios o diseños, los denominados Métodos Mixtos o *Mixed Methods*, los Estudios de Modelos mixtos o *Mixed Model Studies* y los estudios de Método Único o *Monomethod* (Tashakkory y Teddlie, 1998). Como se comentó en la introducción, en el grupo de Métodos Mixtos se engloban aquellos estudios que incorporan metodologías cuantitativas y cualitativas en distintas fases del estudio, mientras que los estudios de Modelos Mixtos aplican ambas metodologías dentro de una misma fase. Por último, los estudios de Método Único implementan distintos procedimientos dentro de un mismo paradigma, ya sea cuantitativo o cualitativo. Dada la relevancia de esta clasificación, situamos los estudios de esta tesis considerando un continuo en el cuál fijamos tres puntos correspondientes a cada tipo de estudio. En uno de los extremos se localizan los estudios de Método Único y en el extremo opuesto los estudios de Métodos Mixtos, dejando en la zona central los Estudios de Modelos Mixtos.

Siguiendo estas indicaciones, los Estudios 1 y 2 formarían parte de los Estudios de Modelos Mixtos ya que, aunque aplican un método cualitativo para obtener la información, los resultados son también presentados de forma “cuantitativa” mostrando tanto porcentajes y frecuencias como relaciones entre categorías y variables por medio de tablas de contingencia. Los Estudios 3 y 5 se clasificarían dentro de la aproximación de Métodos Mixtos porque ambos

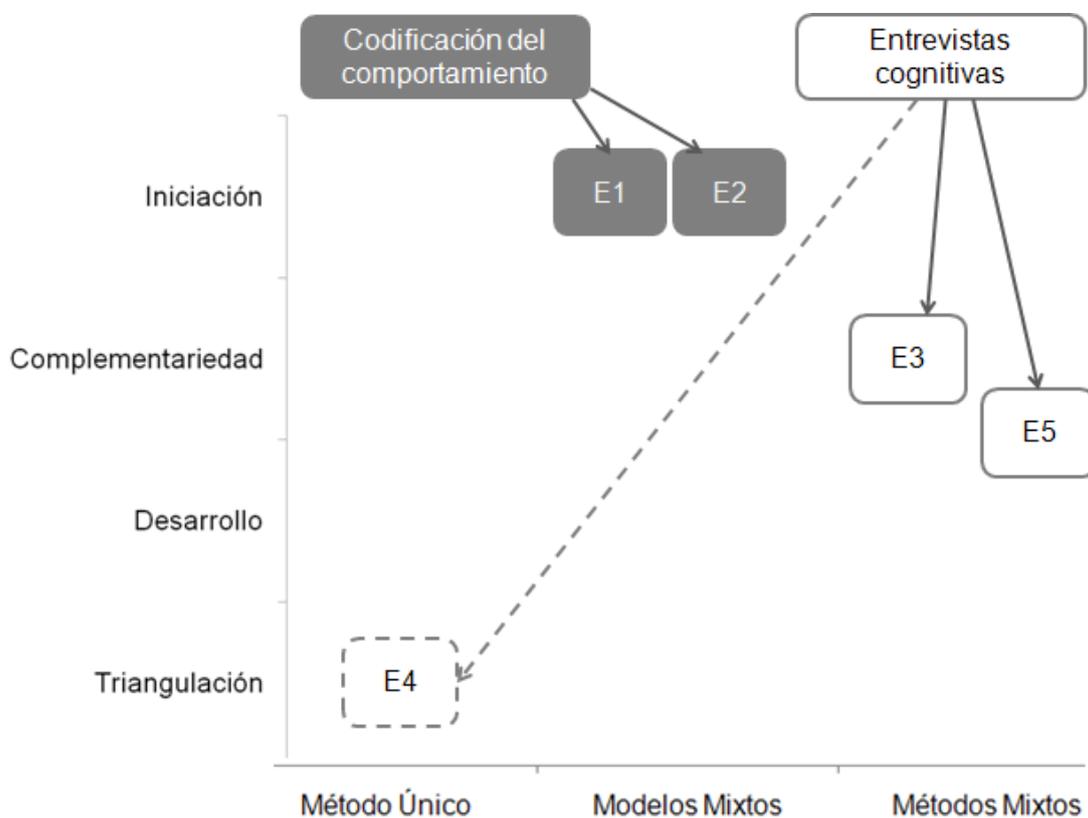
incluyen una etapa cuantitativa y otra etapa cualitativa, es decir, dos mini-estudios en los cuales se utilizan diferentes datos y diferentes métodos para obtener evidencias complementarias. Por último, el Estudio 4 se situaría en el espacio de Método Único ya que se aplicaron dos procedimientos para analizar el DIF y ambos fueron de naturaleza cuantitativa.

Por último, la forma de combinar los resultados puede determinarse teniendo en cuenta la clasificación realizada por Greene, Caracelli y Graham (1989). Estos autores realizan una revisión de los estudios realizados dentro de la investigación mixta y entre otras características describen diferentes formas de combinar los resultados proporcionados por diferentes métodos.

Cuatro son las formas en que se combinan los resultados de nuestros estudios: iniciación, triangulación, desarrollo y complementariedad. La iniciación se refiere a la búsqueda de nuevas perspectivas a partir de los diferentes métodos. En este caso, desde el punto de vista metodológico lo interpretaríamos como la ampliación de las indagaciones posibles o capacidades de un método para obtener información. Los Estudios 1 y 2 responderían a este objetivo porque su principal fin metodológico fue conocer la calidad de la información aportada por la CC en un nuevo contexto, la evaluación de la precisión de las respuestas dadas por los informantes proxies. De esta forma se evaluaron las aportaciones que realizaba la CC para responder a una necesidad nueva, utilizándose los resultados para expandir el rango de posibilidades que ofrece este método. La triangulación, por su parte, se focaliza en la búsqueda de la convergencia y la contrastación de resultados. El Estudio 4, encajaría en este objetivo por aplicar dos procedimientos diferentes para tener mayor seguridad en los datos, es decir, para corroborarlos. Por otra parte, la complementariedad busca clarificar e ilustrar los resultados obtenidos mediante un método a partir de los resultados proporcionados por otro método, es decir, busca enriquecer las conclusiones obtenidas. Esto ocurre en los Estudios 3 y 5 en los que se utilizan las evidencias obtenidas a partir de las EC para interpretar los resultados proporcionados por los procedimientos psicométricos.

Finalmente, el Estudio 5 además de la complementariedad podría situarse en la categoría desarrollo. Esta forma de combinar los datos implica utilizar los resultados de un método para diseñar o detallar otro. En este estudio, los análisis del DIF realizados en el Estudio 4, determinaron las características del diseño de las EC aplicadas en el Estudio 5. La Figura 1, muestra gráficamente la relación entre todas las dimensiones.

Figura 1. Relación entre los estudios, el método cognitivo aplicado, el diseño de investigación y la forma de combinar los resultados.



Como muestra la figura, el eje de abcisas refleja el diseño de investigación, el eje de ordenadas muestra la forma de combinar los resultados y en la parte superior de la gráfica se muestran los métodos de pretest cognitivo empleados. Los estudios se sitúan en el interior de la gráfica respondiendo a su localización en estas tres dimensiones.

INTRODUCTION

The instruments of psychological and sociological evaluation currently form part of our daily lives. Telephone companies do customer satisfaction surveys, hotels evaluate our experience, at airports our travel habits are examined and in schools our performance. Without realizing it we respond to questions in an increasingly automatic manner and do not perceive the importance and implications of such assessments. However, behind all these instruments, there is a long and complex process whose ultimate goal is to ensure that the quality of the information is adequate and meets the proposed needs. Ensuring the quality of information obtained requires the availability of methodologies for assessing the suitability of measurements in different areas such as satisfaction surveys, questionnaires for psychological evaluation, achievement tests, cross-cultural studies, and so on.

The search for methods to improve the quality of measurements has been developed separately in two contexts: survey research and psychological testing. Both contexts have been historically different in important aspects such as the attention given to the different types of error (Groves, 1989; Van de Vijver, 1998). However, currently they coincide in the dissatisfaction with the usual treatment of measurement quality, showing a growing interest in developing procedures to optimize the development of tests and questionnaires and to improve the quality of the information provided.

In the context of survey research this interest is reflected in the renewed concern about measurement errors, which together with sampling error, coverage error and non-response errors are the traditional classifications of the main sources of error given by Groves (1989). Measurement errors include so-called "observational errors" that collect sources of bias related to the measuring instrument, the interviewer, the interviewee and the method of data collection (Groves et al., 2004). These sources of bias reflect the broadening of the points of interest that now look beyond the content of the instrument.

Similarly, in the context of psychological testing, analysis of the quality of the measurements has been carried out by applying psychometric procedures

that have focused on the analysis of the distribution of item responses or on the capacity of these to discriminate between people with different levels of the variable. Nevertheless, over time, emerging needs in the applied field have changed and this change has affected the evolution of the concept of validity itself. For example, the need to compare different demographic groups, linguistic, cultural, etc., has stimulated the design of comparative studies, which in turn has led to further development of quantitative and qualitative procedures used to obtain results on the differences between groups. In turn, the interest in improving the quality of survey studies has increased the attention given to the methods and therefore, has led to development and rapid changes in procedures and the methods of cognitive pretest. This trend has also accommodated new methodological paradigms such as the MR which proposes the combination of quantitative and qualitative procedures. What is clear is that the present goal, in this continuous evolution of methodologies and research needs, is to improve the process of obtaining validity evidences favoring the formulation of appropriate conclusions about the differences between the groups obtained from a broad methodological perspective.

The two-way relationship between the development of methodology and concerns in the applied field, as well as the constant presence of validity as a mediator of relations between the two, is also a defining feature of the studies of this thesis. Therefore, before presenting the work, some relevant points are described that will enable a better contextualization of its content. First the contents related to the Theory of Validity, MR and cognitive pretest methods are presented. Then, the applied problems which are the subject of this research are introduced: the use of indirect or proxy informants and interpretation of psychometric analyses and Differential Item Functioning (DIF). Finally, the objectives of this thesis are stated and the studies are described, taking in to account some dimensions related to methodological features.

The role of validity evidence

'Standards', the short common name of the manual prepared in 1954 by the American Psychological Association, the American Educational Research Association, and the National Council of Measurement in Education, has articulated over successive editions the consensus on the theory and practice in the use of tests, being a reference for tracking the evolution of conceptions about validity in different historical moments. However, this consensus has not been nor is it now exempt from discussion as reflected in the recent monographs and articles on the Theory of Validity (Sireci, 2009; Zumbo, 2009).

Validity has gone from being a requirement considered at the end of the preparation of the test or questionnaire, to permeating the whole test development and evaluation process, while also the ways and means of performing validation studies have increased. Two of the most recent and striking contributions have been the replacement of the categories or types of validity by the notion of "validity evidence" and "sources of validity evidence" along with the appearance of the argument-based approach aimed to guide the validation process proposed by Kane (1992).

Focusing on the last two editions of the *Standards*, the most important arguments of the current framework of validity can be highlighted. The fourth edition of the *Standards* (APA, AERA, & NCME, 1985), already reflected the need to think of validity as a unitary concept, far from the widely criticized preceding division "categories of validity." At that time, the categories of validity were abandoned to start talking about validation strategies aimed at collecting different kinds of evidence. However, this change was not sufficient because the researchers were not satisfied with the role of "construct validity" and demanded further changes in the definition to place the construct in a central position, and also include the social consequences of use of the tests. These ideas were incorporated in the fifth edition of the *Standards* (APA, AERA, and NCME, 1999), a time in which the visibility of validity, according to

Hambleton and Pitoniak (2002), had increased due to the growing use of tests in the critical decision making of people and institutions: recruitment, selection, diagnosis, graduation, etc. The fifth edition of the *Standards*, current version of reference until the publication of the sixth edition scheduled for late 2012 or early 2013, extends validity to all phases of the process of construction and the use of tests, which involves defining validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed scores interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself." (APA, AERA, & NCME, 1999, p. 9). As highlighted in the definition, the key process is to accumulate sufficient evidence to support the proposed interpretation, obtaining this evidence from different sources.

For the first time, the concept of "validity evidence" plays a key role in the validation process. The *Standards* develop this concept of grouping validity sources into five categories:

- a) Evidence based on the content of the test: which comes from the analysis of the relationships between the test content and the construct being measured;
- b) Evidence based on internal structure: which indicate the extent to which relationships between test items and test components conform to the construct on which the interpretations of the test scores are based;
- c) Evidence based on the relationship with other variables: including the analysis of the relationship between test scores and "external" variables. For example, measures of some criterion expected to predict test scores and relationships with other tests that measure the same construct, and tests that measure different constructs and related issues;
- d) Evidence on the consequences of the use of the test" as well as evaluating the appropriateness of the use of the test (Shepard, 1997), the ethical implications associated with the interpretations of the scores and the

social consequences associated with the use of the test (Messick, 1989), and;

- e) Evidence based on response processes", using theoretical and empirical analysis of the respondents' response processes during the test, to obtain evidence on the fit between the construct and the detailed nature of the performance or response actually put into practice by respondents.

The methodologies used should also respond to this grouping, that is, they are determined by the specific source of validity evidence to be addressed. In relation to the evidence based on response processes, the *Standards* specify they be drawn from the analysis of individual responses, so as to ask the respondents about their strategies when responding to questions or about their responses to items, to obtain evidence to enrich the definition of the construct.

Nevertheless, the lack of information on how to obtain evidence based on response processes and how to interpret them caused those researchers uncertainty about how to approach the study of this type of evidence. Currently, as indicated by Zumbo and Shear (2011), remarkable growth is observed in studies focused on obtaining evidence of validity based on response processes, compared to the number of studies conducted on the basis of more traditional sources of evidence. In many of these studies, the collection of evidence is done by applying cognitive interviews or other similar methods.

The *Standards*, when referring to the characteristics of the evidence, refer to the accumulation of both quantitative and qualitative evidence, reinforcing the new trends that combine different types of methodologies to increase the strengths of each and reduce the weaknesses that arise when they are used exclusively. Therefore, MR is presented as a possible paradigm for addressing the search for validity evidence through the combined use of quantitative and qualitative methodologies. Next, we describe the fundamentals of MR that are present in all studies included in this thesis.

Mixed Research

Mixed research (MR) is based on the need to respond to research problems whose complexity is unsatisfactorily addressed by a single method or a single data type. The constraints posed by "methodological" exclusivity raises what is called "pragmatism", a view that prioritizes the importance of the research objective and adjusts the methods to this service. Johnson and Christensen (2008) placed the concept of pragmatism at the core of this paradigm for being what determines that what matters is not whether the researchers consider themselves as quantitative or qualitative, but whether the methods used facilitate the achievement of the objectives that they are pursuing.

Therefore, what promotes MR is the combination of quantitative and qualitative methods in cases in which this combination will bring researchers closer to a better response to the research problem. Most definitions formulated in MR refer to this combination of methods but from different approaches. One of the most relevant definitions may be that stated by Tashakkori and Creswell in 2007 to open a new trend with the first article of the first issue to the Journal of Mixed Methods Research. They define MR as "research in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry" (p. 4).

MR is presented therefore as a new paradigm that seeks to reconcile the traditional debate between quantitative and qualitative research (Reichardt, & Rallis, 1994). Numerous studies conducted in recent years have implemented mixed designs, however, authors such as Johnson and Christensen (2008), consider its legitimization as a "third paradigm" did not come until the publication of the *Handbook of Mixed Methods in Social and Behavioral Research* (Tashakkori, & Teddlie, 2003). These authors published their first book in 1998 which described MR as a paradigm that combines quantitative and qualitative

methods to obtain more informative and sophisticated data (Tashakkori, & Teddlie, 1998).

The influence of MR has grown especially since the 90's, though its beginnings were earlier as shown in the review conducted by Greene, Caracelli and Graham (1989). The authors classified the most common designs depending on the characteristics and objectives of the studies which applied MR. As mentioned above, Tashakkori and Teddlie in 1998 published their first book in this area "Mixed Methodology: Combining the qualitative and quantitative approaches", which was followed by other publications including the book "Handbook of Mixed Methods in Social and Behavioral Research" in 2003 motivated by the growing number of thesis in the field. The second edition of this book appeared in 2010. Moreover, the increase in studies that applied the fundamentals of MR in 2007 led to the creation of the journal: *Journal of Mixed Methods Research*.

The development has been so rapid that the ISI Web of Knowledge contains 2800 papers published that include the term "Mixed Method" as a topic. These works fall into three main areas: "education research", "public health" and "psychology". Focusing on the area of "psychology" it can be seen that the 294 studios in this area were published after 1998, a time that can be considered, also on a theoretical level, the beginning of the most productive period. In addition, there has been an increase in the number of works in recent years, 157 studies (over 50%) were published between 2010 and 2012. As for the appearance of other terms relevant to these works, 18 studies collected from within the keywords the term "*validity*" and eight found the term "*test development*". Moreover, the number of researchers involved in the development of MR increases as shown by the *Mixed Method International Conference* which annually reunites professionals and in 2012 will hold its eighth edition.

Bibliometric data show the importance of a movement in which multiple perspectives are collected, because although all studies classified as MR implement several procedures, the way to do so is very versatile. This

versatility is reflected in the variety of possible designs that have been classified according to different criteria. One of the most widely used design classifications is the one by Creswell (1995) which includes two dimensions: sequentiality and dominance. According to sequentiality, studies can be simultaneous or sequential, i.e., they can apply the two types of methods in parallel or in stages, while dominance describes the priority, i.e. the quantitative may be more dominant than the qualitative or both equally relevant. By combining these two dimensions six types of basic designs are obtained which become nine if the order is considered, a factor that would affect only the sequential studies.

Creswell (1995) also describes a system of representation combining fonts and punctuation. For example, "QUAN + qual" indicates that the study is performed sequentially, the first part to run being the quantitative and the more dominant. "QUAN / QUAL" indicates that both parts were simultaneous and neither was dominant. Considering the characteristics of the study, the previous design (QUAN / QUAL) may be, for example, a study that applies quantitative analysis to analyze multiple-choice questions of a survey and qualitative analysis of the open questions in the same survey. Much more common are studies that include two mini-studies, one using a quantitative perspective on some data and another using a qualitative perspective on other different data, which are later incorporated in order to give combined conclusions.

The situations described are specific examples of what is called Mixed Methods. However, within MR there are other variants, such as the so-called Mixed Model Studies (Tashakkori, & Teddlie, 1998). This category includes studies using quantitative and qualitative methodologies in the same phase, e.g. studies examining qualitative data to develop categories which are then summarized as frequencies or contingency tables. The last group of studies collected under the name of MR are those which, as indicated by Brannen (2005), implement two procedures although both are part of the same paradigm. Tashakkori and Teddlie (1998) call them "Monomethod Studies". In

all three variants, although the design changes, the goal is to access as much information and that this information is of the highest quality.

In the context of MR, needs have been raised related to situations where data from one type of method, such as qualitative, are used to complete data from another method, such as quantitative. Within the qualitative paradigm, cognitive pretest methods are among the most widely used methods in recent years. Their flexibility and ease of use have led to their application being increasingly broad and varied. Among the methods of cognitive pretest, cognitive interviews are currently one of the most known and deployed, leading us to consider their use in different contexts and to ask whether, cognitive pretest methods could be applied to solve some of the problems emerging in the applied field?, for example, could cognitive pretest methods be used to interpret quantitative results?, or could cognitive interviews be used to help explain the "numbers" provided by psychometric tests?

Below is a brief introduction to the cognitive pretest methods in order to facilitate understanding of the methodological approach used to design the studies included in this thesis. It will be shown how cognitive pretest methods can facilitate validation studies within the MR paradigm.

Cognitive pretest methods

Cognitive pretest methods arise from the combination of two factors: dissatisfaction of survey researchers with the treatment routinely given to measurement errors classified as "non-sampling error" by Groves (1989), and the change of paradigm that led to the emergence of cognitive psychology and symbolic interactionism (Foddy, 1996). Both factors prompted the change from the "automatic model", almost as a stimulus-response, to a model that included the behavior of the person answering the questionnaire, ie a more complex

model to explain the "cognitive process" involved in the response process (Willis, 2005).

Undoubtedly, the most cited model for illustrating the new concept of the presence of cognitive processes that occur between the question and answer is the "question-and-answer" cognitive model that characterizes the movement called *Cognitive Aspects of Survey Methodology* (CASM). This movement emerged as a result of two conferences: the *Advanced Research Seminar on Cognitive Aspects of Survey Methodology* held in the United States in 1983 and the *Conference on Social Information Processing and Survey Methodology* which took place in Germany in 1984 (Jabine, Straf, Tanur, & Tourangeau, 1984). Both conferences stressed the importance of taking people into account as "active agents" and considering the cognitive processes involved in the response process. Tourangeau (1984) further developed the "question-and-answer" process model that included, in addition to the traditional elements, a description of the cognitive processes involved in the "question-and-answer" process. Figure 1 shows a model in which the cognitive processes that appear between the formulation of the question and the statement of the response are positioned.

Figure 1. Representation of the question-and-answer model

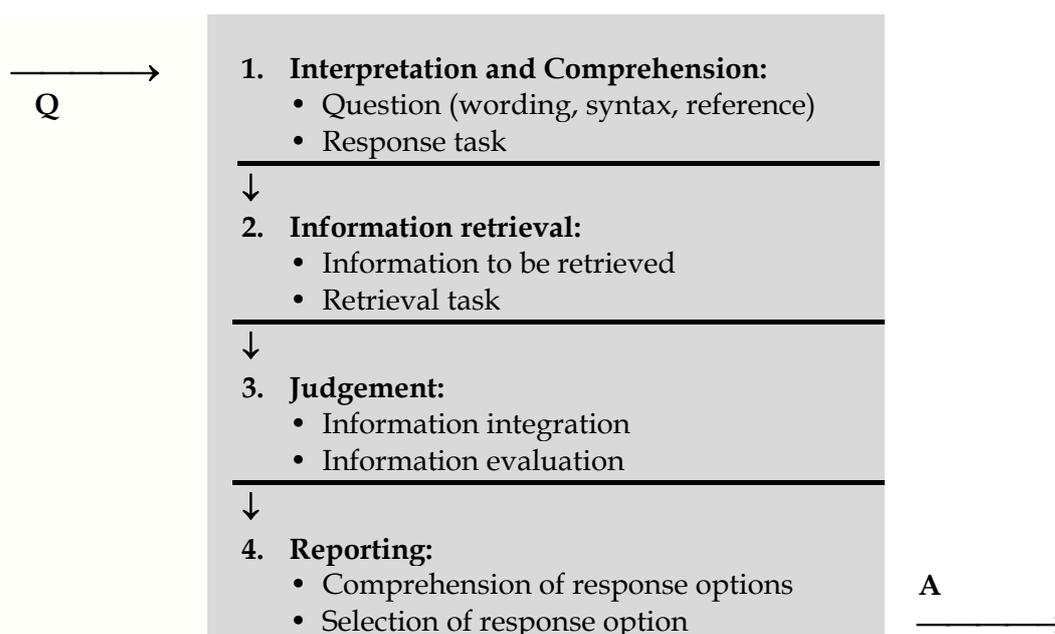


Figure 1 shows the four phases that a person would complete when responding to a question. During these four phases that occur between the question reading and the answer communication, the participants develop several cognitive operations: first they interpret and understand the question or task ahead which involves understanding both the intended purpose and the concepts and expressions that are included, and then they retrieve the information needed to answer the question, then make a judgment that allows them to integrate and evaluate the information retrieved, and finally they adjust their response to the proposed alternatives and communicate it.

In the last two decades, the model has evolved to include the possible non-sequentiality of the phases in all circumstances (Collins, 2003), and the presence of social and cultural dimensions. The social and cultural dimensions are key to explaining the motivation of both the interviewer and the interviewee in the "question-and-answer" process. For example, one question may lead to different responses depending on whether it is formulated in a formal or informal context. Krosnick (1999), noted the importance of motivational aspects emphasizing the possible influence of processes called "optimization" and "satisfaction". These processes relate to the motivations, whether intrinsic or extrinsic, that each participant has when answering a question. Optimization requires the performance of all the phases of the process and to do so completely. The alternative to optimization is satisfaction. A respondent "satisfies" when he jumps some stage of the "question-and-answer" process or completes it superficially. Satisfaction appears in those participants who want the interviewer-respondent interaction to end in a "satisfactory" way for both. Factors such as the difficulty of the question, the education level of respondents, the presence or absence of rewards, etc., have been shown to have a consistent influence on levels of "satisfaction".

The aim of cognitive pretest procedures is to trace the "question-and-answer" process to obtain evidence for two purposes: "to optimize" the questions and infer what they are "really" measuring.

Some of the cognitive pretest methods most used are Cognitive Interviews, Behavioral Coding, debriefing or expert appraisal (Presser, et al. 2004). This introduction will focus on the description of the first two, as they are used in the studies included in this thesis.

Cognitive Interviews

Cognitive Interviews (CI) are the cognitive pretest method most used in studies seeking to understand the process of "question-and-answer" carried out by the participants (Willis, 2005). The success of this method is partly due to its flexibility and ability to adapt to the research objectives. Use of CI has been conducted mainly in the official statistical institutes, which have followed their own approach to cognitive psychology. However, it is necessary to mention another more sociological approach to CI in which the mediating presence of social and cultural factors is considered, giving the interviewee a more focused role in translating his life experiences to the "question-and-answer" process (Miller, Chepp, Willson, & Padilla, 2012, in press).

Among the many definitions, which arise from the cognitive perspective, of what is a CI, one of the most comprehensive and agreed upon definitions is that stated by Beatty and Willis (2007), who describe it as "the administration of a draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (p. 288). The definition calls for a prior determination of the intentions of the researcher, which is referred to an interview protocol to assess the degree of fit between the intended construct and information obtained.

The interview protocol can be developed following a *think-aloud* method (Ericsson, & Simon, 1980) focused on the verbalization of the thoughts of participants as they respond to objective questions, or following the *probing*

based method (Willis, DeMaio, & Harris-Kojetin , 1999), which develops follow-up probes for specific areas of each question. In the latter case, the probes are developed based on the features or elements of the questions that researchers consider potentially problematic. For example, when it is suspected that the meaning of a word can be confusing a follow-up probe can be developed in order to reveal the interpretation made by the respondents. The information obtained on the different possible interpretations will allow adjustments to be made to facilitate the transmission of the intended concept. Table 1 shows six types of follow-up probes specified by Willis (2005), and an example to illustrate the purpose of each.

Table 1. Types and examples of follow-up probe

Follow-up probe	Example
General probe	How did you arrive at that answer? Tell me what you were thinking
Comprehension/ Interpretation probe	What does the term "health" mean to you?
Paraphrasing	Can you repeat the question I just asked in your own words?
Confidence judgment	How sure are you that you went to the doctor five times in the past 12 months?
Recall probe	How do you remember that you went to the doctor five times in the past 12 month

There are also different approaches to the analysis of data from CI. On one hand, there are models based on coding schemes. These models summarize the results of CI using usually between two and nine general categories (Willis, DeMaio, & Harris-Kojetin, 1999). The categories are developed based on the characteristics and objectives of the investigation, although a common tendency is to propose a grouping that responds to what the researcher was "looking for", which is then supplemented by what the researcher "has found" without

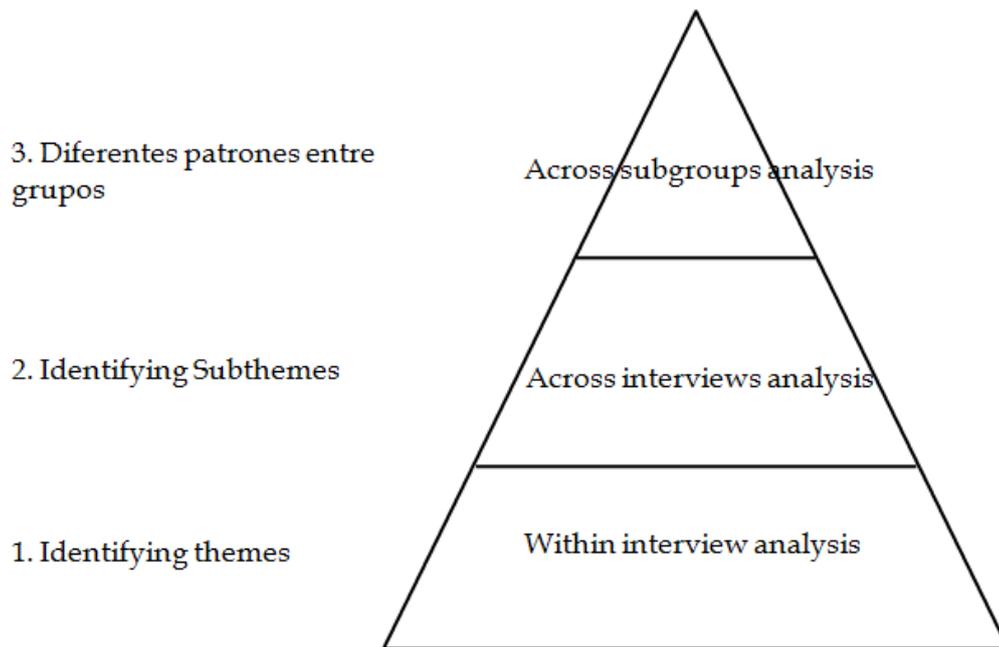
looking. For example, Willis (2005) proposes five general categories to guide a first review of the results. These categories are:

- a) Item specific recommendations for changes to wording;
- b) Need for further Specification of objectives or the manner in which the question satisfy them;
- c) Problems related to the ordering and other interactions between survey questions;
- d) Problems related to reduction in overall instrument length or burden, and;
- e) Limitations on what can be asked of survey respondents using the intended procedures.

Among the models based on coding schemes, a more detailed analysis may be proposed by specifying the categories listed. To do this, firstly a review would be done of the narratives that emerged before the protocol. Next, possible objective categories would be determined based on key elements of these narratives and these categories would be allocated to the segments of narratives. For example, segments derived from the protocol could be classified by grouping what happened during the interview depending on the stage to which it belongs within the "question-and-answer" model. That is, segments of narrative could be divided according to the interpretation that the participant has made of the concepts included in the question or the processes involved in adjusting their response to the proposed alternatives.

From the sociological approach, strategies for qualitative analysis have been developed focused on the discourse and narratives of the interviewees. For example, Miller (2007) developed a model that includes three stages in the analysis of CI. Figure 2 shows the graphical representation of this latest model incorporating some modifications on the original proposal (Miller, 2007).

Figure 2. Graphical representation of the pyramid model



As shown in Figure 2, the three phases of analysis range from the most elementary aspects linked to the data, up to the most complex located at the top. First, the individual analysis of the interviews allows for *themes*. The issues relate to general themes developed by participants. For example, if a person is asked what they were thinking about when answering a question about daily activities, that person may list places, situations, people, etc. These general categories would be the themes. The aim of the second level is to compare the interviews to obtain *sub-themes*. The sub-themes are specifications of the themes. For example, in the above case we could divide the theme "people" in "family" and "not family" or "family", "friends" and "co-workers" depending on our goal. Finally, the third level seeks to compare different groups in order to detect different patterns of interpretation and related to certain characteristics of the participants. In the example above, a result typical of the third level could be finding that women most frequently used the sub-theme of "family" and men the sub-theme "co-workers."

The contribution of this model is the differentiation of the findings in these three levels that can raise different kinds of conclusions, i.e. that the results taken at each level of the pyramid can be used to achieve different objectives. The results of phase 1 can for example guide an amendment or correction of some questions so as to ensure the fit of the respondents' interpretation to that of the researcher, while the results of phase 3 can be used to establish differences in the interpretation of constructs across different population groups. But, could these results be used to interpret differences between groups detected by other methods such as an analysis of differential item functioning (DIF)? Could they be used to enrich and complement the results obtained using psychometric methods? These are some of the questions that have arisen in the studies in this thesis.

Behavior coding

The Behavior Coding (BC) developed by Cannell in the 60's (Cannell, Fowler, & Marquis, 1968), focuses on another aspect considered fundamental for locating problematic elements or characteristics of the questions: the interviewer-respondent interaction. BC is a procedure based on the systematic observation of respondent. It involves coding the events during the interaction that takes place between the interviewer and the respondent when a questionnaire is administered. To perform the encoding different categories of observation may be included according to the interests of the researchers. In addition, there are coding schemes that summarize the key aspects that should be recorded about the interviewer and respondent. One of these is the coding scheme of Oksenberg, Cannell and Kalton (1991), which describes different categories defined by the actor (interviewer or respondent), and the moment (during the reading of the question or during the response). The combination of the categories leads to various codes that reflect the most relevant behaviors

occurring during the interviewer-respondent interaction. Table 2 shows some of the behaviors included in the standard classifications, and a brief description.

Table 2. Categories for the classification of the behaviors occurring during interviewer-respondent interaction

Actor	Moment	Codes	Description
Interviewer	Question reading	Exact	Interviewer reads the question exactly as printed
		Slight changes	Interviewer reads the question changing a minor word that does not alter question meaning
		Major changes	Interviewer change the question such that the meaning is altered or does not complete reading the question
Respondent	During the answer	Interruption	The respondent stops the question reading (to request clarification or to answer)
		Request repetition	Respondent asks for repetition of question
		Request clarification	Explicit expression for indicating problems in the comprehension of the concepts included in the question or in the task comprehension
	During the answer elaboration	Changes answer	Respondent changes the answer
		Expresses doubts	Respondent expresses doubts during the answer elaboration
		Mismatch answer	The response is adequate but is not exactly worded as any of the answer options
		Invalid answer	The response is not related to the question
		Don't know answer	The respondent does not know how to respond

Actor	Moment	Codes	Description
		Qualified answer	The response indicates uncertainty.
		Adequate answer	The response fits the objective of the question

The behaviors listed in Table 2 are divided firstly according to the person who performs them: interviewer or respondent. The interviewer's most important behaviors occur during the reading of the question, since this is the moment when the interviewer can make modifications to the established script. However, the relevant behaviors of the respondent can occur both during the answer and during the answer elaboration. The participant's response usually occurs during the last phase, which is also coded according to the degree to which it fits the intended objectives.

Depending on the behaviors occurred during the answer and during the answer elaboration, the sequences can be classified. A sequence is everything that happens from the presentation of one question to the presentation of the next question. Following the instructions in Ongena and Dijkstra (2006) there are three types of sequences. paradigmatic sequences are those considered "ideal", ie where the delivery of the question is identical to that indicated in the interview protocol, the respondent answer is adequate and the interviewer recognize it as such (exact reading + adequate answer).

Deviations from this "ideal" sequence are known as non-paradigmatic sequences, and these in turn can be classified as problematic or non-problematic. The first group includes those that have negative effects on the data, such as the interviewer giving a reading with large changes or the respondent providing an inappropriate answer. Among the non-problematic are those sequences in which something problematic has occurred but it has been resolved or does not adversely affect the data. An example of this would be when the participant asked to repeat the question because some unexpected noise occurred that prevented him from hearing the statement.

Traditionally, the results obtained through BC have been used to locate potential problematic areas during the interaction and thus point out the aspects that should be paid more attention. For example, the results of BC may show the need to include the definition of a concept when the participants have often asked for clarification of a question. But, could BC be used in other contexts? For example, could BC be used in a more interpretive manner, to assess the quality of the answers of respondents with specific characteristics? Could it also provide other information such as the convergence between the responses of different types of informants providing a more precise data offered by other methods? Some of the studies included in this thesis show data relevant to these proposals.

Once situated in the methodological framework, from the general to the more specific, we address the research problems in the case studies of this thesis to illustrate the reasons that have led to the development of each one.

Proxy respondents

We begin by placing the issue of indirect or proxy informants in the context of survey research, where typically the questionnaire being applied is aimed at a specific population, ie, it is designed with the people who will respond in mind. The problem is that sometimes it is not possible to locate the self-reporter, or the survey's "target" person at the address, or that person can not answer due to their health condition or legal restrictions. This poses difficulties in terms of organization and costs, resulting in the need to seek solutions. One possible solution is to ask another person (proxy), who is present in the home, to answer questions by putting himself in the place of the person for whom the questionnaire was intended (self-reporter). In this way, researchers avoid having to come back another time, replacing the target person, or losing any information of a participant and thereby increasing the rate of non-response.

The obvious consequence of the decision to interview a proxy is that the information obtained is not the direct information that was sought, so possible discrepancies between different sources of information should be considered. The use of "proxies" has been widely criticized by researchers, which has led the Eurostat Task Force to limit the use of proxies to situations that are strictly necessary such as self-reporter being unable to answer for health or legal reasons (Eurostat Task Force, 2005; Tafforeau, Lopez, Tolonen, Scheidt-Nave, & Tinto, 2006).

Nevertheless, the situations in which there is no access to the self-reporter are still present and the usefulness of proxies remains clear in other everyday settings outside those legal or personal limitations. One of these scenarios may come from the frequent incompatibility of interviewer-respondent times perhaps for the length or characteristics of the working day of one of them or both. The fact that "needing" the use of proxies causes the researchers responsible for the survey administration to consider investigating the actual quality of the responses provided by these informants as well as the convergence between these responses and those of the self-reporter. The evaluation of the convergence between the responses of self-reporters and proxies, and the evaluation of the quality of the responses of proxies have been traditionally studied in parallel. The quality of the responses of proxies is considered adequate when it coincides with that given by the self-reporter, ie, the self-reporter's answer is considered the "gold standard".

On analyzing the information obtained in studies using proxies, it is found that it mainly evaluates the convergence of responses based solely on statistical indicators and ignores the qualitative aspects (Pickard, Johnson, Feeny, Ashfaq, Carriere, & Abdul, 2004, Todorov, & Kirchner, 2000). When this assessment is made exclusively from statistical results, it can cause the loss of relevant information related to the content of the answers, as well as a comparison of incompatible data. For example, using only statistical indices requires transforming the answers and coding them to compare them "statistically". This

coding is usually done by assigning numbers to response categories. When following this procedure, situations such as no answer or "I do not know" answer are coded as a null response, for example, with a zero.

But, what happens when a null response is compared with any other answer? In this case we have a situation of disagreement that would be numerically represented the same way as another situation in which two people gave valid responses but different. That is, when a person answers "yes" and another "no", we face a situation of disagreement that until now has been equivalent to another situation in which a person responds "yes" and the other says "I do not know". This threat to the comparability is due to how often what happened has been categorized or labeled as "agreement" or "disagreement" between informants, rather than analyzing the nature of the disagreement. However, in some contexts it may be relevant to differentiate between the types of disagreements. In fact, differentiation is seen as essential when we want to locate, for example, questions that a group of respondents have not answered consistently, while there has been a response from another group. On the other hand, it may be important to distinguish differences between valid responses when you are interested in detecting inconsistencies between informants in a question, e.g. "Yes/No". Therefore, we need more than the label of "disagreement". This information, that may be valuable, is not obtained with the methods currently applied for assessing convergence, however, it could be accessible if substantive categories are added to the usual coding of responses to give us more information about what is happening.

Another problem in research with proxies is selecting them. When an interviewer arrives at a home looking for his self-reporter and does not find them, the most efficient and easiest way is to use any person present at that time as a proxy. However, in addition to the usual controlled sociodemographic variables, aspects not normally considered such as the type of relationship between the proxy and self-reporter can affect the quality of data, i.e. the proxy being a self-reporter's brother is not the same as being his roommate. Also

relevant are aspects such as the content of the question. For example, it was observed that the proxies give appropriate responses when asked about observable health areas and shared experience with others, such as problems with daily physical activities or chronic limitations (Magaziner, Bassett, Hebel, Maldini, & Gruber, 1996), while their responses are more limited when asking about the emotional state, pain and other health dimensions that are not directly observable (Grootendorst, Feeny, & Furlong, 1997).

In short, it suggests several situations or needs that could be met if we manage to improve the process of assessing the quality of the responses of proxies and the analysis of the convergence between the responses of different informants. This also could help to shed light on the characteristics required by the questions designed to be answered by proxies or even the characteristics of the "perfect proxy"; elements that, as indicated by Rajmil, et al. (1999) are, together with the relationship between the informants, the main factors influencing the responses of proxies.

The improvement of the process involves, on the one hand, avoiding assessments based solely on quantitative results, supplementing this data with qualitative information that may be relevant and, on the other, recognizing nonequivalent comparisons between the responses of the participants. This brings us to the issues that have been addressed in the studies of this thesis: is it possible to assess the quality of the responses of proxies by themselves, that is, separate from the response of self-reporter, although this is the standard comparison?, i.e., can we know if the proxy responses are accurate even though we do not know the answers of the self-reporter? Does the type of relationship between the informants have any effect on the comparison of the agreement? Moreover, could this assessment include qualitative information that will help us discover the source of disagreement between informants? Could the analysis of the behavior of participants be distilled so that the kind of "disagreement" between informants could be revealed?

Psychometric analysis and Differential Item Functioning (DIF).

We come now to the last point we must address in order to frame all the pieces in this thesis. In this final section, some aspects of psychometric analysis and Differential Item Functioning (DIF) are introduced that will help in understanding the aim of the study proposed in the last works of the thesis.

We are so familiar with psychometric tests that sometimes we routinely apply them to obtain information about the quality of the instruments administered. Some of the most widely used statistical indices in all the disciplines are the mean and variance, which describes the distribution of responses and their variability, or the discrimination index (corrected item-total correlation), which quantifies the ability of an item to differentiate between the levels of the variable that people show. Also coefficient alpha or factor analysis are widely used procedures for estimating reliability and describing the internal structure of the instrument. These analyses allow us to identify general strengths and weaknesses of the questionnaire or test we are using. Nevertheless, sometimes unexpected results are obtained that we do not know how to interpret or whose causes cannot be explained. The same occurs in the analysis of DIF, although in this case the implications are greater because it can affect the conclusions, undermining the validity of interpretations and/or hurting one group over another.

DIF occurs when examinees with the same proficiency level on the characteristic or attribute measured, but who belong to different groups (i.e., demographic, linguistic, national or cultural), have a different probability of giving a specific item response (Millsap and Everson, 1993). When we analyze the DIF we compare two groups that have traditionally been referred to as Reference Group (RG) and Focal Group (FG). The designation of the groups can be performed using traditional criteria, such as considering the larger group or the one that responds to the original version of the instrument as the RG, or responding to specific study objectives, such as using the target group as RG.

According to the distribution of DIF, it may be uniform or non uniform among the groups depending on whether there is interaction between the measured attribute and group membership (Mellenbergh, 1982). There are a variety of statistical techniques to evaluate DIF, whose functioning has been extensively studied in recent years (Zumbo, 2007, Hidalgo, & Gomez-Benito, 2010). Among them, the Mantel-Haenzsel (MH) and Logistic Regression (LR) are the most used procedures and those that were applied in Studio 4 of this thesis which are described in more detail.

Previous sections discussed the importance of existing methodologies being able to adapt and respond to the research problems that arise in each historical period. DIF is a good example of this, since its rise is due in part to the effort to meet the needs that arise in the context of cross-demographic and/or cultural studies. The fact that more and more situations occur where researchers need to compare groups makes DIF analysis gain importance and its development is addressed in more depth. DIF can identify those problematic areas that, being present in the instrument we are using, can prejudice a group by making us reach non real conclusions. DIF detection locates characteristics of the item that, because they operate differently in the groups involved, can cause differences unrelated to actual differences in measured attribute. Thus, DIF warns us about that we should take the precaution when drawing conclusions and about the problems that we must solve to be confident about the equivalence between the groups.

Because of the concern that this threat to the validity of comparisons between groups has sparked, the number of studies that apply DIF analysis for assessing instruments has grown significantly in recent years (Cho, Martin, Conger, & Widaman, 2010; Kalaycioglu, & Berberoglu, 2011; Lewis, Yang, Jacobs, & Fitchett, 2012). However, the assessment of what is causing the DIF has not had such successful results. Despite the research of the causes of DIF has attempted to address this on different fronts, for example through the study of the characteristics of the instruments or items, or through the evaluation of

the quality of adaptation, still no conclusive results have been achieved. Therefore, in recent times, the study of the causes of DIF has been approached from novel perspectives and through the application among others of qualitative techniques or the use of multilevel models (Swanson, Clauser, Case, Nungester, & Featherman, 2002; Van den Noortgate, De Boeck, & Meulders, 2003).

Although efforts have been increasingly intense, we have not yet managed to discover the causes of DIF and their origin. Our proposal consists of applying cognitive pretest methods to assist in the investigation of those causes. We propose evaluating the utility of these methods to help us interpret these unexpected results obtained from psychometric analysis and that can not be explained. We also raise the possibility of using CI to explain and interpret the results of DIF and to clarify the search for causes.

Having developed both methodological and conceptual issues addressed in this thesis to a very basic level, we now outline the objectives and structure of the included studies.

OBJECTIVES AND STRUCTURE

Specifying what was seen in the introduction, it is clear that one of the pillars of this thesis is the convergence between the qualitative and quantitative methods to assess the quality of measurements in different contexts. The goal that has characterized the process during which this project was carried out has been to address methodological problems associated with the quality of the information provided by questionnaires and scales, to obtain validity evidence by applying cognitive pretest methods and quantitative methods, especially, psychometric. To achieve this overall objective, the following specific objectives were proposed, to be achieved using the individual studies:

- Specific objective 1: To determine the ability of Behavior Coding (BC) to evaluate the responses of proxies. Studies 1 and 2.
- Specific objective 2: To evaluate the usefulness of Cognitive Interviewing (CI) as a procedure for interpreting psychometric results. Studies 3 and 5.
- Specific Objective 3: Determine the ability of CI to locate the causes of DIF. Studies 4 and 5.

In turn, the studies carried out to achieve these objectives can be described using three dimensions: the applied cognitive method, the research design and how to combine the results. We situate each of the studies taking this information in to account.

First, in terms of the cognitive method used, in two studies BC was applied (Study 1 and Study 2) and in two others CI (Study 3 and Study 5). In the first two studies, BC allowed us to evaluate the accuracy of the answers given by proxies through the analysis of interviewer-respondent interaction. In Study 1 the quality of the answers given by proxies was assessed, while in Study 2 convergence between the responses of proxies and self-reporters was observed. Moreover, in Studies 3 and 5 CI were applied to obtain information on the interpretations made by the participants when responding to the items of a scale. Study 4 did not directly apply any of the methods of cognitive pretest, however, the design of this study was conducted with the aim of obtaining

information that would enable the development of the follow-up probes for the CI protocol used in Study 5. In turn, the information provided by CI led to interpreting the data from the study that analyzed the DIF, so it could be placed within the second group of studies.

The second dimension concerns the research design. Following the traditional division of paradigms in quantitative and qualitative, mixed research emerges, and within it are described three types of studies or designs, so-called Mixed Methods, Mixed Model Studies and the Monomethod Studies (Tashakkory, & Teddlie, 1998). As discussed in the introduction, in the group of Mixed Methods those studies incorporating quantitative and qualitative methodologies in different phases of the study are included, while Mixed Models Studies applied both methodologies within a single phase. Finally, Monomethod Studies implemented various procedures within the same paradigm, either quantitative or qualitative. Given the importance of this classification, we place the studies in this thesis considering a continuum in which we fix three points for each type of study. At one end are located Monomethod Studies and at the opposite end Mixed Methods Studies, leaving Mixed Model Studies in the centre.

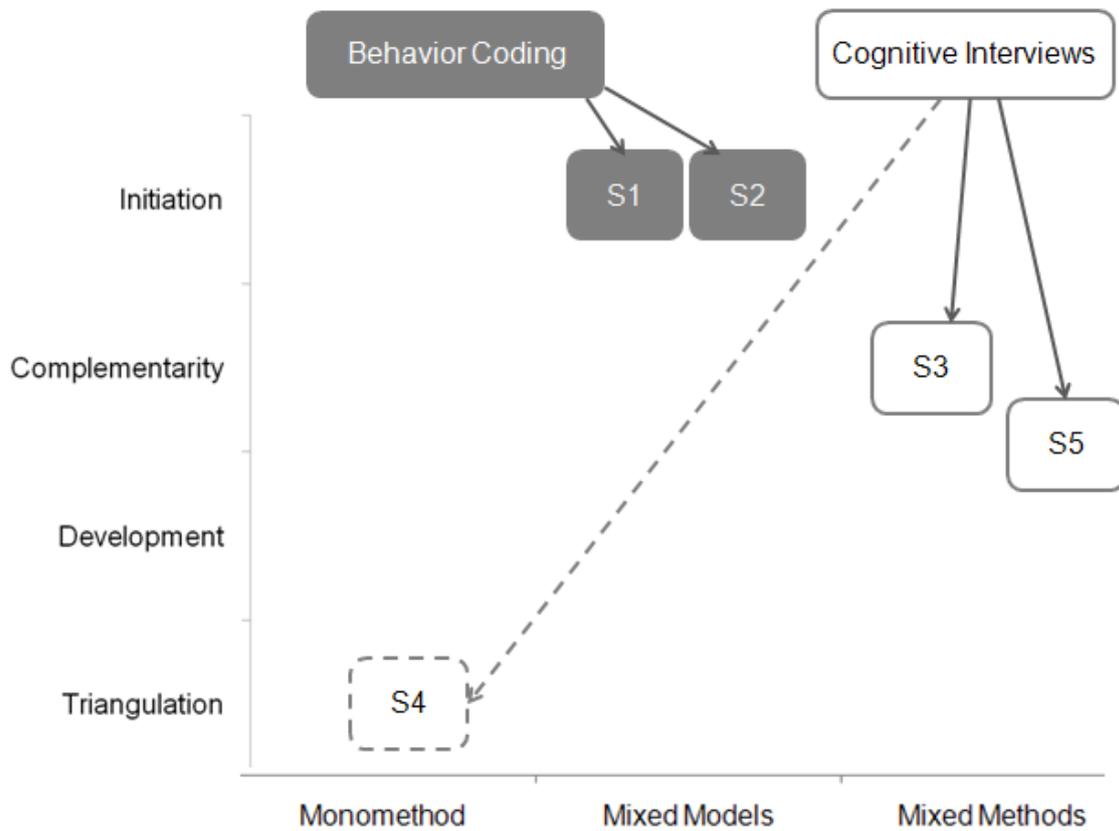
Following these indications, Studies 1 and 2 form part of the Mixed Models Studies because while applying a qualitative method for obtaining the information, the results are also presented in a "quantitative" format showing both percentages and frequencies as relations between categories and variables using contingency tables. Studies 3 and 5 would be classified within the Mixed Method approach because both include a quantitative and a qualitative stage, ie, two mini-studies which used different data and different methods to obtain additional evidence. Finally, Study 4 would be located in the space of Monomethod as two procedures were applied to analyze the DIF, and both were quantitative in nature.

Lastly, the way of combining the results can be determined taking into account the classification made by Greene, Caracelli and Graham (1989). These

authors performed a review of studies conducted in MR and among other features they describe different ways of combining the results provided by different methods.

There are four ways to combine the results of these studies: initiation, triangulation, development and complementarity. Initiation refers to the search for new perspectives based on the different methods. In this case, from the methodological point of view we would interpret it as the extension of possible inquiries or capacity of a method for obtaining information. Studies 1 and 2 respond to this objective because its main methodological purpose was to determine the quality of the information provided by the BC in a new context, the evaluation of the accuracy of the answers given by proxy informants. In this way the contributions being provided by the BC to meet a new need were evaluated, using the results to expand the range of possibilities offered by this method. Triangulation, meanwhile, focuses on the search for convergence and contrasting results. Study 4, would fit this objective by applying two different procedures for greater data reliability, ie to corroborate. Moreover, the complementarity seeks to clarify and illustrate the results obtained by a method based on the results provided by another method, that is, it aims to enrich the findings. This occurs in Studies 3 and 5 in which the evidence obtained from CI is used to interpret the results provided by psychometric procedures. Finally, Study 5 in addition to complementarity could be placed in the development category. This way of combining data involves using the results of a method to design or elaborate another. In this study, the DIF analyzes conducted in Study 4, determined the design features of the applied CI in Study 5. Figure 1 graphically shows the relationship between all dimensions.

Figure 1. Relationship between the studies, the cognitive method applied, the research design and the method of combining the results.



As shown, the abscissa axis reflects the research design, the vertical axis shows the method of combining the results and the top of the graph shows the cognitive pretest methods employed. The studies are located inside the graph in response to its location in these three dimensions.

CHAPTER 1:

Analysis of Quality of Proxy Questions in Health Surveys by Behavior Coding

Benítez, I., Padilla, J.L., y Ongena, Y. (2010). Analysis of Quality of Proxy Questions in Health Surveys by Behavior Coding. *Western Journal of Nursing Research*, XX(X), 1-16. On line first. DOI:10.1177/0193945910388049 (wjn.sagepub.com)

Abstract

The aim of this study is to show how to analyze the quality of questions for proxy informants by means of behavior coding. Proxy questions can undermine survey data quality due to the fact that proxies respond to questions on behalf of other people. Behavior coding can improve questions by analyzing interviewer-respondent interactions. 29 proxies participated in the pre-testing of a disability questionnaire. The questionnaire includes 11 questions related to daily life limitations as a result of health problems. Interviewer-proxy interactions were coded and analyzed by means of the Sequence Viewer program. The percentages, of from methodological perspective, of ideal “question-and-answer” sequences varied from 28% to 76% throughout the 11 questions analyzed. The results obtained pointed out the necessity of reviewing some of the proxy questions analyzed. Behavior coding can improve the quality of proxy questions in health surveys when proxy informants are surveyed.

Keywords: Behavior coding; proxy informants; health surveys.

Introduction

The use of indirect informants or "proxies" to obtain information about other household members is common in household health surveys (Duncan, et al., 2002; Magaziner, et al., 1996; Pickard, et al. 2004; Schwarz, & Wellens, 1997). A "proxy" is a person who answers survey questions about the health conditions of other people, whereas "self-reporters" answer about themselves. Proxy reporters are often employed to fill in the designated household rosters in household surveys. The answers of the proxy reporter determine the eligibility of other household members to respond to other sections or questionnaires used in the survey. In addition, the use of proxies is frequent in medical processes or disease evaluations. Proxies have been used in an evaluation of the quality of life for reporting on patients with communication difficulties due to cerebral injuries (Sneeuw, et al., 1997). In turn, proxy and self-reporter answers have been compared to evaluate the validity of a questionnaire for patients who have suffered a stroke (Teixeira-Salmela, Devaraj, & Olney, 2007).

The guidelines and quality criterion for the design of health surveys prepared by the Eurostat Task Force, summing up the present consensus about the use of proxy reporters, indicates that the use of proxy-reporters should be limited only to cases in which: a) people are incapable of responding to questions, due to serious health problems (e.g. dementia, physical or severe mental disability, etc.); or b) to those for whom it is not possible to interview for legal reasons (i.e. minors) (Tourangeau, 2003). Nevertheless, using proxies is a common practice in national statistical institutes for the accomplishment of health surveys in numerous countries. The *Health Examination Survey (HES)* database (Koponen & Aromaa, 2001) promoted by the *Scientific Institute of Public Health (SIPH)* includes information provided by 34 countries from surveys in which proxy informants have been used, among them: Belgium (*Health Interview Survey*), Czech Republic (*Labour Force Sample Survey*), France (*Survey on Household Living Conditions*), Holland (*Continuous Quality of Life Survey*), and in Spain (*Survey of Disability, Personal Autonomy and Dependence*

Situations). Proxy informants were also used in the *National Health Interview Survey on Disability* for the National Center for Health Statistics in United States (Todorov, & Kirchner, 2000).

Few studies have looked at how to increase the quality of the answers provided by proxies, in spite of the fact that the use of proxies has been traditionally considered a threat to the quality of survey data (Ávila-Funes, Gray-Donald, & Payette, 2006). One of these studies evaluated the bias in the proxy answers by means of the *National Health Interview Survey on Disability* carried out in New York. The results showed that proxies used different response strategies than self-reporters (Todorov, & Kirchner, 2000). Another study, in which the proxy answers were evaluated on a scale of cerebral injury impact, showed that proxy and patient evaluations are more consistent when they evaluate observable and specific behaviors, whereas the agreement decreased when the proxy informants made subjective judgments (Duncan, et al., 2002). On the other hand, while evaluating the quality of life in patients who have suffered cerebral injuries, it was found that the proxies' evaluations were sensitive to the differences in the patients' functionality (Sneeuw, et al., 1997).

Evaluating proxy responses is especially challenging in disability survey contexts because, according to the WHO (World Health Organization, 2009) definition of disability, the classification of a person as having or not having a disability is a subjective judgment as it depends on the interaction between social conventions, individuals, cultural norms, expectations, etc. Therefore, the responses to the questions about whether or not a person has a disability could vary according to the type of informant (self-reporter versus proxy), as a result of potential differences between both norms and expectations, but not necessarily as a result of objective information.

Pretest methods can be helpful in improving survey questions. The general objective of pretest methods is the identification of the causes of errors in surveys by means of the analysis of the events occurring during the "question-and-answer" process (Willis, 2005). Behavior coding is one of the pretest

methods used by survey methodologists, either on its own or in combination with other pretest methods such as cognitive interviewing, focus groups or speech analyses, to optimize the question drafting and the questionnaire design (Presser, et al., 2004). In contrast with such pretest methods, behavior coding provides systematic, objective, and replicable results (Groves, et al., 2004).

The behavior coding method was developed in the 1960's by Charles Cannell to evaluate both the questions and the interviewer behavior (Cannell, Fowler, & Marquis, 1968). Behavior coding is based on the rationale that the interviewer's and respondent's behaviors provide information about potential problems with survey questions related to question phrasing and to questionnaire design by systematically observing the interviewer-respondent interaction (Blair, & Srinath, 2008). Moreover, behavior coding allows survey researchers to evaluate the quality of survey questions aimed at specific respondent groups defined by characteristics such as "age", "educational level" or "gender". Nevertheless, little attention has been given to questions designed for respondents with different roles (self-reporter or proxy) in the interview process.

The aim of this study is to show how to analyze the quality of proxy questions by means of behavior coding in a health survey. In this study the adequacy of the questions to be answered for proxies will be also discussed.

Method

Participants

29 proxy informants, 13 men and 16 women with an age average of 31.06 years, took part in the pretest of a disability questionnaire. The educational level of participants was balanced (14 participants with less than 14 years of schooling and 15 participants with more than 14 years of schooling). The sample size of the study is within the interval (15-50) recommended by several authors to maximize the usefulness of results provide by the behavior coding method (Blair, & Srinath, 2008).

All participants were Spanish and they provided information only about people whom they lived with and had a direct familiar relationship with, for instance, parents, partners, brothers or sisters. The selection was carried out with regard to various requirements that determine if the participant was "eligible", that is to say, they had the same characteristics of the target population of the future health survey in which the tested questions in this study would be administrated.

It was also confirmed that the participants had not previously taken part in a survey pretest. The participants were contacted via associations for disabled person support and they received 30 Euros for taking part in the study.

Materials

The people responsible for carrying out the interviews used interview protocols during the pretest which included demographic questions and 11 "target" questions. The "target" questions were the selected questions to be analyzed during the pretest by means of behavior coding. These questions were selected by experts who evaluated the questions of the questionnaire, identifying those questions which could present difficulties. Experts had a long experience in the field of health surveys and survey methodology. Table 1 shows the 11 questions to be analyzed by means of the behavior coding method.

Table 1. Selected questions from the disability questionnaire.

Target questions
Q. 1. Is there any person in your home who has been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year.
Q.2. Is there any person in your home who has serious difficulty speaking in an understandable manner and saying meaningful phrases without help?
Q.3. Is there any person in your home who has serious difficulty understanding the meaning of what others say without help?
Q.4. Is there any person in your home who has serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille.
Q.5. As a result of problems of a cognitive or intellectual nature, is there any person in your home who has serious difficulty when intentionally using the senses? For example, paying visual attention, listening attentively, etc.
Q.6. As a result of problems of a cognitive or intellectual nature, is there any person in your home who has serious difficulty learning to read, write, count (or calculate), copy or difficulty learning to use everyday utensils?
Q.9. Is there any person in your home who has serious difficulty showing other people affection, respect or transmitting feelings including physical contact such as kisses, caresses, etc.?
Q.10. Is there any person in your home who has serious difficulty forming and maintaining family relationships?
Q.11. Is there any person in your home who has serious difficulty forming and maintaining sentimental or sexual relationships with a partner?

Procedure

The interviews, in which the questionnaire with the target questions was applied, were conducted by two trained and experienced interviewers (one male and one female). They were specifically instructed to ask target questions as the questions were worded in the questionnaire. The interviews were conducted in a laboratory specially equipped to perform cognitive pre-testing. Confidentiality and the exclusive use of the information for research purposes were assured. Having obtained the respondents consent, the interviews were audio and video recorded. The interviews were transcribed and two coders used the transcripts and recordings to systematically classify the interviewer and respondent behaviors. The two coders worked independently and once first classifications were made, they met to analyze discrepancies and reach an agreement.

Verbal behavior coding

The behavior coding was done by means of the Sequence Viewer program (Dijkstra, 2008). Coders were also trained by experts in Sequence Viewer program. This program provides information about possible problems with the content or the format of the questionnaire, by systematic classification of behaviors occurring during the interview. The analysis begins with the division of the transcripts into sequences. A sequence starts with the reading of a question and ends when the reading of the following question starts (Dijkstra 1999). The sequences are analyzed by assigning different codes depending on the behaviors occurring during the interviewer-respondent interaction. For example, while interviewers are asking questions, respondents can ask for explanations or extra information (coded as "request for clarification"), and respondents can interrupt the interviewer giving their answers to the question before the interviewer has finished reading or making comments (coded as "interruption"). Answers given by the respondent after interviewers have finished reading the question can be classified in different ways, of which the classification realized by Oksenberg, Cannell and Kalton (1991) is the most

commonly used. This classification has been extended by authors like Van der Zouwen and Smit (2004), Forsyth, Levin and Fisher (1999), and Ongena (2005). Table 2 shows the coding scheme used in this study, which is primarily based on the classification by Oksenberg, Cannell and Kalton (1991).

Table 2. Categories for the classification of respondents' behaviors.

Codes	Meaning
<i>During the question reading</i>	
Request clarification	Explicit expression for indicating problems in the comprehension of the concepts included in the question or in the task comprehension.
Interruption	The respondent stops the question reading (to request clarification or to answer).
<i>Answer</i>	
Mismatch answer	The response is adequate but is not exactly worded as any of the answer options
Invalid answer	The response is not related to the question
Don't know answer	The respondent does not know how to respond
Qualified answer	The response indicates uncertainty
Adequate answer	The response fits the objective of the question

In order to evaluate the quality of proxy questions, codes were used in the study as indicators of response accuracy. A scale of accuracy was developed, using extremes represented by the codes "adequate answers" (being the most accurate) and "invalid answers" (being the most inaccurate). The intermediate categories were defined as "mismatch answer", "qualified answer" and "don't know answer".

Depending on the combination of codes assigned, sequences are classified as: "paradigmatic sequences", "non paradigmatic-non problematic sequences" and "non paradigmatic- problematic sequences". A "paradigmatic sequence" is

defined as the ideal sequence during the question-and-answer process. An ideal sequence is that in which the delivery of the question is identical to that indicated in the interview protocol, the respondent's answer is adequate and the interviewer recognizes the answer as being adequate (Ongena & Dijkstra, 2006). A “non paradigmatic sequence” is problematic or non problematic depending on whether the type of behavior occurring is considered to be a problematic influence on the data. In this study, the occurrence of “mismatch answers”, “invalid answers”, “don’t know answers”, “qualified answers” and “requests for clarification” all classify the sequence as a problematic sequence. A sequence is classified as non paradigmatic-non problematic when deviations occur that are not problematic (for example, interruptions).

Sequences are classified considering the codes assigned to each behavior occurred during the sequence. For example, the occurrence of the behavior “request clarification” causes a sequence become non paradigmatic although the respondent’s answers was adequate.

Once the sequences were classified, a frequency analysis was performed which consisted first of calculating the frequencies of each type of sequence followed by calculating the rate of the occurrence of problematic answers. When 15% or more of a question’s administrations show one or more problematic interactions is a widely accepted criterion for determining if a question is flawed (Blair & Srinath, 2008). On the other hand, if the percentage of non paradigmatic sequences is considered, questions in which the percentage is greater than 60% must be checked (Van der Zouwen & Dijkstra, 2002). Analysis of 319 sequences (i.e., 11 questions x 29 respondents) was conducted utilizing both criteria to illustrate the use of behavior coding in the study.

Results

For the analysis 319 sequences (i.e., 11 questions x 29 respondents) were taken into account.

Types of sequence

First, the behavior coding analyses showed the frequency of the occurrence of each type of sequence produced by the proxy informants. Table 3 shows the percentages of occurrence of each type of sequence for each target question.

Table 3. Frequencies and percentages of each type of sequence.

Target questions	Type of sequence					
	Paradigmatic sequence		Non paradigmatic- non problematic sequence		Non paradigmatic - problematic sequence	
	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.
Q.1.Habitual activities	28	8	21	6	52	15
Q.2. Speak	76	22	3	1	21	6
Q.3. Understand	66	19	24	7	10	3
Q.4. Use the phone	41	12	17	5	41	12
Q.5. Use the senses	66	19	10	3	24	7
Q.6. Learn	69	20	17	5	14	4
Q.7. Move the body	76	22	10	3	14	4
Q.8.Change posture	66	19	14	4	21	6
Q.9. Show affection	76	22	14	4	10	3
Q.10.Family relationships	72	21	10	3	17	5
Q.11.Sentimental relationships	55	16	34	10	10	3

As Table 3 shows, the percentage of paradigmatic sequences, that is to say, ideal sequences from the methodological point of view, ranges between 28 % and 76 % for the target questions. The Cramer's V statistic value (.2762) indicates a low association between the type of sequence and the target question analyzed. Target question 1 showed the highest percentage of non paradigmatic-problematic sequences (52 %). Following the usual criteria, question 1 was recommended for checking, because 72 % of the sequences were

classified as non paradigmatic. This high percentage could be due to the content of the question, which is more general and ambiguous than the rest of the target questions.

Codes for proxy responses

In this study we were particularly interested in deviations produced by proxies. Table 4 shows the percentages of adequate and (four types of) inadequate answers for the 11 questions of the disability questionnaire. The percentages per row add up to more than 100% since multiple behaviors can occur in one sequence. For example, the respondent can change the answer after “invalid answer” code.

Table 4. Frequencies and percentages of answer category codes.

Target questions	Codes									
	Mismatch answer		Invalid answer		Don't know answer		Qualified answer		Adequate answer	
	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.	Perc.	Freq.
Q.1. Habitual activities	45	13	0	0	0	0	6	2	90	26
Q.2. Speak	0	0	6	2	0	0	3	1	90	26
Q.3. Understand	0	0	7	2	0	0	0	0	97	28
Q.4. Use phone	14	4	21	6	3	1	7	2	76	22
Q.5. Use senses	7	2	7	2	3	1	7	2	93	27
Q.6. Learn	10	3	3	1	0	0	3	1	100	29
Q.7. Move the body	10	3	3	1	0	0	0	0	100	29
Q.8. Change posture	10	3	10	3	3	1	0	0	90	26
Q.9. Show affection	3	1	7	2	0	0	3	1	93	27
Q.10. Family relationships	3	1	3	1	0	0	0	0	97	28
Q.11. Sentimental relationships	3	1	3	1	3	1	10	3	90	26
Cramer's V	.403		.209		.149		.180		.239	

As table 4 shows, the Cramer's V values reveal a low association between the type of answer produced by the participants and the question analyzed in all the cases except for the code "mismatch answer". This code shows the highest percentage of occurrence for the set of target questions. Question 1 achieved the highest percentage of "mismatch answers" (45 %). The following example represents a situation in which a "mismatch answer" was produced:

Interviewer: Is there any person in your home who has been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year. Yes, seriously limited; yes, limited but no seriously; not.

Respondent: "Yes"

The answer given by the respondent was coded as a "mismatch answer" because it does not fit to any of the response alternatives offered. The high percentage of mismatch answers found in question 1 might be due to respondents understanding it as a "yes / no" question without considering the three response alternatives offered.

Question 4 achieved the highest percentage of "invalid answers" (21 %), and the lowest percentage of "adequate answers" (76%). The following example represents an "invalid answer" found in question 4:

Interviewer: "Is there any person in your home who has serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille".

Respondent: "In my home nobody knows how to use the machines for writing in Braille".

The answer was coded as an "invalid answer" because its content is not related to the intended objective of the question.

Finally, in question 11, 10 % of answers were registered as qualified. An example from the interviews illustrates the meaning of the "qualified answer" code.

Interviewer: "Is there any person in your home who has serious difficulty forming and maintaining sentimental or sexual relationships with a partner?"

Respondent: "I don't think so".

In question 11, the high percentage of qualified answers may indicate that the proxies have doubts when responding to questions on personal topics such as sexual or personal relationships. Nevertheless, this leaves the interviewer with a dilemma; should she further probe for an unqualified answer, or just accept the answer as given. In some cases this is not necessary, as respondents may spontaneously repair their qualified answer by giving an unqualified adequate answer afterwards.

Difficulty indicators when asking questions

"Request clarification" and "interruption" are codes commonly used in behavior coding as indicators to identify difficulties while interviewers are asking questions. Table 5 shows the frequencies of both codes in the 11 target questions.

Table 5. Frequencies and percentages of difficulty indicator codes during the question reading.

Target questions	Codes			
	Request clarification		Interruption	
	Perc.	Freq.	Perc.	Freq.
Q.1. Habitual activities	7	2	3	1
Q.2. Speak	3	1	3	1
Q.3. Understand	0	0	3	1
Q.4. Use the phone	14	4	0	0
Q.5. Use the senses	3	1	0	0
Q.6. Learn	3	1	0	0
Q.7. Move the body	0	0	0	0
Q.8. Change posture	7	2	0	0
Q.9. Show affection	3	1	0	0
Q.10. Family relationships	7	2	0	0
Q.11. Sentimental relationships	3	1	0	0
Cramer's V	.1740		.1591	

As Table 5 shows, the Cramer's V values reflect a low association between the behaviors produced by the participants and the target question analyzed.

Requests for clarification occurred most frequently with question 4. Questions 8 and 10 also showed a high percentage in the appearance of this code. The following example demonstrates the occurrence of such a “request clarification”:

Interviewer: *“Is there any person in your home who has serious difficulty forming and maintaining family relationships?”*

Respondent: *“family relationships?”*

Interruptions were coded to some extent in question 2. The excerpt illustrates an interruption found in question 2:

Interviewer: *“Is there any person in your home who has serious difficulty speaking...”*

Respondent: *“Yes”*

Interviewer: *“... in an understandable manner and saying meaningful phrases without help?”*

In this specific case, difficulties could arise from an interruption, since the respondent is answering the question before hearing all the elements that have to be considered (Van der Zouwen & Dijkstra, 2002).

Discussion

The aim of the study was to illustrate how to analyze the quality of the questions intended for proxy respondents in a health survey by means of behavior coding. The results from the behavior coding application to the disability questionnaire pretested in the study allowed the quality of the proxy questions to be analyzed.

The general results showed percentages of paradigmatic sequences between 28% and 76 % for the set of 11 target questions. Only question 1 "habitual activities" achieved more than 60 % of non paradigmatic sequences.

The results highlighted some questions to be checked or in which it was necessary to examine the proxies' behavior in detail. These problems might be due to the characteristics of the questions, or to the role represented by the informants. For example, question 1 ("habitual activities") is worded as a "yes / no" question while three alternatives are offered to the respondent. This is a problem that is common in survey questionnaire design (Ongena 2003). In addition, two of the three options are "positive" ("Yes, seriously limited" and "yes, limited but no seriously), and one is "negative" ("not"). Thus, researchers find an "adequate answer" in cases in which the respondents' answer is negative, but a large percentage of "mismatch answers" when the respondents' answers are positive but they replied with a simple "yes". Assessing how serious the limitation was and distinguishing between the affirmative alternatives can be a difficult task for proxies. On other hand, proxy behavior could cause measurement error because either proxies focus on aspects which are not the aim of the question (question 4 on the use of the telephone), or they face non-observable or sensitive topics (question 10 on "family relationships"). Possible impact of demographic such as "educational level", "degree of family relationship", and so on, on proxy questions were not specifically addressed in our study due to its particular design which can be considered a limitation.

When using proxies, survey researchers consider several factors. The difficulty of the task and the motivation for responding to questions could be different for self-reporter than for proxies. In addition, proxy respondents may have less information available in their episodic memory (Schwarz & Wellens 1997). More studies focused on comparing the proxies and self-reporter behavior are necessary, as well as evaluating the convergence between the answers provided by both type of informants. Future researches may address these topics.

Respondent behavior can be studied from multiple perspectives, including more qualitatively oriented studies. For example, Collins, Shattell, and Thomas (2005) address how to deal with potentially problematic interviewee behaviors,

such as flattery, filtration or statements indicative of social desirability response bias for qualitative research. Behavior coding as a method, provides a systematic approach to analyze interviewer and respondent behavior, is flexible, and offers the possibility of obtaining qualitative and quantitative information which help survey methodologists improve survey data quality. In comparison with other pretest methods, behavior coding is focused in the participant's behavior. The assumption behind behavior coding is that the interviewer-respondent interaction can provide very useful information about potential problems with question phrasing and questionnaire design. This information allows survey researchers to identify questions with high percentage of "problematic behaviors" as questions which should be revised.

Behavior coding also presents some limitations. For example, it is possible a respondent gives an adequate answer although he has not understood the real sense of the question. In fact, there may be a "gap" between respondents' "observed" behaviors and their understanding of the key concepts in the questions. Combining behavior coding and cognitive interviewing can resolve that "gap". Future researches in the pre-test methods field should address how to combine evidence provide by different pre-test methods.

On the other hand, it is necessary to reach a greater consensus about the criteria used to check the questions based on the results obtained by means of the behavior coding. In a review of the studies in which behavior coding is used, it was found that some authors consider those questions in which the percentage of adequate answers was lower than 85% to be problematic, whereas others authors think questions must be checked when the percentage of adequate answers is lower than 90%; while others focused on the percentage of inadequate answers, recommending to review the questions in which the percentage is greater than 15% (Van der Zouwen & Smit, 2004). The criteria used can cause changes in the conclusions obtained because, for example, an adequate answer can occur after an inadequate answer. If an inadequate answer

criteria is used a question can be eliminated although a high percentage of final adequate answers has been reached.

Behavior coding has shown its usefulness for evaluating the quality of the questions designed for proxy informants by providing detailed information about the participants' behavior and facilitating the detection of possible sources of measurement error. However, more research is needed to find out the causes of question problems identified by coding behavior and their consequences when results of behavior coding studies are applied in survey questionnaire design, especially when proxy questions are included in the survey questionnaire. Nevertheless, as Oksenberg, Cannell, & Kalton (1991) highlight, there is convincing evidence of the usefulness of behavior coding to improve the quality of survey questions providing quantitative, systematic, and replicable results.

References

- Ávila-Funes, J.A., Gray-Donald, K., & Payette, H (2006). Medición de las capacidades físicas de adultos mayores de Quebec: un análisis secundario del estudio NuAge. *Salud Pública en México*, 48, 446-454.
- Blair, J., & Srinath, K.P. (2008). A note on sample size for behavior coding pretests. *Field methods*, 20 (1), 85-95.
- Cannell, C.F., F.J. Fowler, & K.H. Marquis. (1968) "The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews". *Vital and Health Statistics, Series 2, No. 26*.
- Collins, M., Shatell, M., & Thomas, S. P. (2005). Problematic Interviewee Behaviors in Qualitative Research. *Western Journal of Nursing Research*, 27, 2, 188-199.
- Dijkstra, W. (1999). A New Method for Studying Verbal Interactions in Survey Interviews. *Journal of Official Statistics*, 15, 67-85.
- Dijkstra, W. (2008). *Sequence Viewer* (version 4.4a). Free University of Amsterdam, Netherlands.
- Duncan, P., Min Lai, S., Tyler, D., Perera, S., Reker, D.M., & Studenski, S. (2002). Evaluation of Proxy Responses to the Stroke Impact Scale. *Stroke*, 33, 2593-2599.
- Forsyth, B., Levin, K., & Fisher, S. (1999). *Test of an appraisal method for establishment survey questionnaires*. Proceeding of the ASA Section on Survey Research Methods. Alexandria, VA: American Statistical Association.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley.
- Koponen, P., & Aromaa, A. (2001). Health Examination Surveys (HES) in the European Union: Review of literature and inventory of surveys in the

- EU/EFTA Member States. Scientific Institute of Public Health 2001. [Retrieved January 13, 2010, from: <http://www.iph.fgov.be/keywords.asp?Lang=EN&ReportID=2107>"].
- Magaziner, J., Speaar, S., Hebel, J.R. & Gruber-Baldini, A. (1996). Use of “proxies” to measure health and functional status in epidemiologic studies of community-dwelling women. *American Journal of Epidemiology*, 143 (3), 283-292.
- Ongena, Y. P. (2003). Pre-testing the ESS-questionnaire using interaction analysis., *European Social Survey CCT meeting, Sociaal en Cultureel Planbureau*. Den Haag: http://naticent02.uuhost.uk.uu.net/questionnaire/pre_testing_interaction_analysis.doc.
- Ongena, Y.P. (2005). *Interviewer and Respondent Interaction in Survey Interviews*. Unplubished doctoral dissertation. Amsterdam Vrije Universiteit.
- Ongena, Y.P., & W. Dijkstra (2006). Methods of Behavior Coding of Survey Interviews. *Journal of Official Statistics* 22, 419-451.
- Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics* 7 (3), 349-365.
- Pickard, A.S., Johnson, J.A., Feeny, D.H., Ashfaq, M.D., Carriere, K.C., & Abdul M.N. (2004) Agreement Between Patient and Proxy Assessments of Health-Related Quality of Life After Stroke Using the EQ-5D and Health Utilities Index. *Stroke*, 35, 607-612.
- Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. New York: Wiley-Interscience.
- Schwarz, N, & T. Wellens (1997) Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Acotrs and Observers. *Journal of Official Statistics*, 13 (2), 159-174.

- Sneeuw, K.C.A., Aaronson, N.K., de Haan, R.J., & Limburg, M. (1997). Assessing quality of life after stroke: the value and limitations of proxy ratings. *Stroke* 28, 1541-1549.
- Teixeira-Salmela, L.F., Devaraj, R., & Olney, S.J. (2007). Validation of the human activity profile in stroke: a comparison of observed, proxy and self-reported scores. *Disability Rehabilitation*, 29 (19), 1518-1524.
- Todorov, A., & Kirchner, C. (2000). Bias in "proxies" Reports of Disability: Data From the National Health Interview Survey on Disability. *American Journal of Public Health*, 90 (8), 1248-1253.
- Tourangeau, R. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research*, 15, 3-7.
- Van der Zouwen, J., & Dijkstra, W. (2002). Testing questionnaires using interaction coding. In D. Maynard, H. Houtkoop-Steenstra, N. Schaeffer, and J. Van der Zouwen (Eds.), *Standardization and Tacit knowledge: Interaction and Practice in the Survey interview*. (pp.427-447). New York: Wiley.
- Van der Zouwen, J., & Smit, J. H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and transcripts of Question-Answer Sequences: A Diagnostic Approach. In S. Presser, J. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp.109-130). New York: Wiley.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.
- World Health Organization (2009). *ICIDH-2: International Classification of Functioning, Disability and Health (ICF)*. Geneva, Switzerland: Author.

CHAPTER 2:

Evaluation of the convergence between "self-reporters" and "proxies" in a disability questionnaire by means of Behavior Coding method

Benítez, I., Padilla, J.L., y Ongena, Y. (2012). Evaluation of the convergence between "self-reporters" and "proxies" in a disability questionnaire by means of behavior coding method. *Quality & Quantity*, 46 (4), 1311-1322.

Abstract

Household surveys often require including proxy reporters to obtain information about other household members who cannot be interviewed. The participation of proxies can undermine survey data quality due to the fact that proxies must respond to questions thinking about other people. The objectives of the present study were to analyze the Behavior of proxy reporters and evaluate the convergence between the answers given by proxies and self-reporters by means of behavior coding. This improves the evaluation of convergence, since only adequate (i.e., interpretable) answers given by both types of informant are taken into account. Responses to a disability questionnaire employed by an official statistical institute were analyzed. The questionnaire includes 11 questions about different limitations related to everyday activities. 16 self-reporter and 16 proxies formed 16 couples whose members lived together and supported a direct family relation. The results show a high percentage (52%) of convergence between both types of informant, although fluctuating across the questions and the couples. Proxies showed relatively more adequate behavior during the interaction than self-reporters. From this we conclude that proxies can be considered at least as good informants as self-reporters from an interviewer-respondent interaction perspective. Future research should address the impact of proxy responses on survey validity.

Keywords: Proxies, behavior coding, convergence evaluation, disability questionnaire

Introduction

One of the most notable characteristics of the design of many household surveys is the use of proxy reporters to obtain information about other household members. A "proxy" is a person who answers the survey thinking about another person of their household or environment, whereas a "self-reporter" answers thinking about himself. Proxy reporters are often employed to fill in the designated household rosters in household surveys. The answers of the proxy reporter determine the eligibility of the other household members for responding to other sections or questionnaires used in the survey. The decision to use proxy reporters is the result of weighing up costs, sampling errors and response errors (Sunghee, Mathiowetz, & Tourangeau, 2007). Nonetheless, the reduction of costs can be outweighed by the increase in measurement errors when compared with self-response reporting.

The use of proxies is a common practice in household surveys carried out by official statistical institutes, despite the fact that the document on guidelines and quality criterion for the design of health surveys prepared by the *Eurostat Task Force* for the design of health surveys, summarizing the consensus among survey researchers, discourages interviewing proxies. The consensus among survey researchers indicates that the use of "proxies" should be limited only to 'replacing' people who are incapable of responding to questions, due to serious health problems (e.g. dementia, physical or severe mental disability, etc.) or those for whom it is not possible to interview for legal reasons, for example minors (Tafforeau, et al., 2006).

Many studies have investigated the influence on survey data quality of the proxies' characteristics, such as age, gender, educational level, level of income, and the relationship with the self-reporter. For example, Magaziner, et al. (1996) found a high degree of agreement between proxies and self-reporters who live together. That study, which included differences in the type of information requested, shows that proxies were able to accurately report on health and observable functioning, such as physical or daily tasks, chronic conditions, etc.

Nevertheless, the information provided by the proxies about the symptoms of health (frequently not observable and not discussed with others) was less precise. Some investigations also found a decrease in the precision when proxies report about psychosocial characteristics or symptoms (Pickard, et al., 2004) or subjective areas like memory and thought, communication, emotion or behavior (Duncan, et al., 2002).

Not much is known about the effects of using proxies on data quality, but it is known that the answers sometimes differ from the answers provided by self-reporters. Pickard, et al. (2004) evaluated the agreement between proxies and self-reporters by means of a questionnaire which evaluates the "quality-of-life" construct, finding systematic differences between the information given by both types of informant. In relation to the accuracy of the answers provided by each type of informant, some studies show that self-reporter' answers are more precise than proxies' answers (Loftus, et al., 1992).

Schwarz and Wellens (1997) showed by means of several experiments that proxy reports show higher consistency than self-reporters. However, consistency does not necessarily mean more accuracy, as the information source for proxies may be biased. Schwarz and Wellens argue that proxies derive information to judge an answer from dispositional information (i.e., the personality and likes and dislikes of the person they are reporting on), whereas self-reporters are more likely to base their judgment on situational factors. Hence, questions on distant events and concerning lengthy reference periods will increase convergence of proxy and self-reporters, since for such questions self-reporters have less possibility of accessing episodic information on situational influences, and consequently, like the proxies, will use dispositional information,.

Nevertheless, the use of proxies is necessary in surveys in which the self-reporter "cannot" be interviewed. This is often the case with surveys about health and well-being, or when respondents have different health conditions

associated with age, for example a study in which proxies were used to retrieve the functional state of patients over 65 years old (Magaziner, et al., 1996).

Evaluating proxy responses is especially challenging in disability survey contexts because, according to the OMS (2009) definition of disability, the classification of a person as having or not having a disability is a subjective judgment as it depends on the interaction between social conventions, individuals, cultural norms, expectations, etc, that is to say, the level of disability is determined by the environment and its demands and not only by the diagnosed difficulties the person has. Therefore, the responses to the questions about whether or not a person has a disability could vary according to the type of informant (self-reporter versus proxy), as a result of differences between both in norms or expectations, but not necessarily as a result of objective information.

Few studies have been focused on the potential sources of errors associated with the “role” assigned to the respondent. Todorov and Kirchner (2000) found a systematic evaluation bias in the proxies’ responses to the National Health Interview Survey on Disability. In this survey proxies were used in the cases when not all household members were available, in order to avoid having to return to the same households on repeated occasions.

Pretest methods, whose usefulness to optimize the information obtained by surveys have been widely proven, can be used to evaluate the influence of proxies on survey data quality. The general objective of pretest methods is the identification of the causes of errors in surveys by means of the analysis of the events happening during the “questions-and-answer” process (Tourangeau, 2003). Tourangeau, Rips, and Rasinski (2000) formulated the most disseminated version of the “questions-and-answer” model with four sequential main phases: comprehension, retrieval, judgement and response selection. The extent to which respondents “pass” through each of these phases could be determined by the role assigned to the respondent: proxy versus self-reporter. Thus, pretest

methods could contribute to detecting differences in the cognitive process completed by proxies and self-reporters.

Among pretest methods, the behavior coding method has proven its utility for providing information about the problems which can exist in relation to the formulation of the questions and the questionnaire format by means of the systematic observation of the interviewer-respondent interaction (DeMaio, Rothgeb, & Hess, 1998). Behavior coding can detect problematic behaviors by classifying the events occurring during the interaction. In a general sense, behavior coding allows the researcher to establish relations between the problematic behaviors identified and the respondent, interviewer and questionnaire characteristics.

Given the potential effects of the use of proxy reporters on measurement error in surveys, it is necessary to evaluate the convergence between proxies and self-reporters. A high convergence between both types of informant would lead to a higher confidence in using proxies when it is not possible to access self-reporting responses. Applying behavior coding methods can improve the evaluation of convergence by analyzing only "adequate" answers, i.e. answers that are directly interpretable as an answer, given by both types of informant. The aim of the present study is to analyze the behavior of proxy reporters and evaluate the convergence between proxies and self-reporters in a disability questionnaire by means of behavior coding.

Method

Participants

Sixteen couples, that is to say, 32 people (13 men and 19 women) participated in the cognitive pretest of the disabilities questionnaire included in a survey. The members of each couple were living together and they had a direct family relation. The selection was carried out with regard to various requirements that determine if the participant is "eligible" for a future administration of the survey. In all cases, the participants should have mastered

functional Spanish, i.e., sufficient to manage everyday situations. With respect to demographic variables, the participants' ages were between 16 and 80 years old. Lastly, it was checked that the participants had not previously taken part in a survey pretest. Table 1 presents the distribution of the demographic variables used for selecting the participants for the cognitive pretest.

Table 1. Description of characteristics of the cognitive pretest participants

Subgroup	Gender			Age (Average)		
	Male	Female	16-25	26-45	46-60	+ 61
Self-reporters	6	10	1 (19)	4 (40.3)	8 (50.6)	3 (66)
Proxy reporters	7	9	6 (20.5)	6 (34.7)	4 (52)	0 (0)

Materials

Two versions of a disability questionnaire that differed in question wording depending on the respondent type, were used. The self-reporter version questions were addressed to the self-reporter with 'you', whereas in the proxy version questions, this 'you' was replaced by 'any person in your home who'. Table 2 shows the 11 questions selected to be analyzed in the questionnaire pretest, called "target questions", in the self-reporters version.

Table 2. Self-reporter target question

Target Questions
1. Have you been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year.
2. Do you have serious difficulty speaking in an understandable manner and pronouncing meaningful phrases without help?
3. Do you have serious difficulty understanding the meaning of what others say without help?
4. Do you have serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille.
5. As a result of problems of a cognitive or intellectual nature, do you have serious difficulty when intentionally using the senses? For example, paying visual attention, listening attentively, etc.
6. As a result of problems of a cognitive or intellectual nature, do you have serious difficulty learning to read, write, count (or calculate), copy or difficulty learning to use everyday utensils?
7. Do you have serious difficulty moving your body from one place to another without changing position, without help and without supervision? For example, going from sitting on the bed to sitting on a chair.
8. Do you have serious difficulty changing posture without help and without supervision? For example, getting up from a chair, lying down on the bed, kneeling down, etc. Exclude the action of moving one's body posed in the previous question.
9. Do you have serious difficulty showing other people affection, respect or transmitting feelings including physical contact such as kisses, caresses, etc.?
10. Do you have serious difficulty forming and maintaining family relationships?
11. Do you have serious difficulty forming and maintaining sentimental or sexual relationships with a partner?

The interviewers used an interview protocol for performing the interviews. The interview protocol included the target questions together with the usual demographic questions. Interviews were recorded in video and audio, having previously obtained the respondents consent.

Procedure

During the recruitment phase, 16 people who met the necessary requirements to act as self-reporters were selected. These people were selected to be self-reporters regardless of whether or not they had any limitations when performing everyday activities. In addition, these people were requested to come to the interviews along with another household member, who would act as a proxy. The participants did not know in advance what their roles would be. Interviews were conducted individually and took place in cognitive laboratories equipped with video and audio recorders. Later behavior coding was carried out using the transcripts and recordings of the interviews.

Analysis

The analysis was done using the program Sequence Viewer version 4.4.a (Dijkstra, 2008). This program provides information about possible problems with the content or the format of the questionnaire, by classifying the behaviors occurring during the interview. The classification of the behavior can be carried out depending on when it occurs: while the interviewers were asking the questions, or while the respondents were answering the questions.

While interviewers are asking questions, respondents can ask for explanations or extra information (coded as “request for clarification”), and respondents can interrupt the interviewer giving their answers to the question before the interviewer has finished reading or making comments (coded as “interruption”). Answers given by the respondent after interviewers finish reading the question can be classified in different ways, of which the classification realized by Oksenberg, Cannell and Kalton (1991) is the most commonly used. This classification has been extended by authors like Van der Zouwen and Smit (2004), Forsyth, Levin and Fisher (1999), and Ongena (2005). Table 3 shows the version of the classification by Oksenberg, Cannell and Kalton used in this study.

Table 3. Responses categories to classify respondents' answers.

	Codes	Meaning
Problematic answers	Mismatch answer	The response is adequate but doesn't coincide with any of the answer options
	Invalid answer	The response is not related to the question
	Don't know answer	The respondent did not know how to respond
	Qualified answer	The response indicates uncertainty
Non problematic answer	Adequate answer	The response fits the objective of the question

Depending on the behaviors occurring, the sequence can be classified as paradigmatic, non paradigmatic-non problematic or non paradigmatic-problematic. A "paradigmatic sequence" is defined as the ideal sequence during the question-and-answer process (Schaeffer, & Maynard 1996). In agreement with Ongena and Dijkstra (2006) an ideal sequence is that in which the delivery of the question is identical to that indicated in the script, the respondent's answer is adequate and the interviewer recognizes the answer as being adequate. A "non paradigmatic sequence" is problematic or non problematic depending on whether the type of behavior occurring is considered to be a problematic influence on the data. In this study, the occurrence of mismatch answers, invalid answers, don't know answers, qualified answers and requests for clarification all classify the sequence as a problematic sequence. A sequence is classified as non paradigmatic- non problematic when deviations occur that are not problematic (for example, interruptions).

After this classification, a frequency analysis was performed which consisted firstly of calculating the frequencies of each type of sequence and secondly of the rate of the occurrence of problematic answers. This analysis

provided information about questions with possible difficulties. Then, an evaluation of the convergence was done to obtain information about the agreement in the answers given by both types of informant. To carry out the convergence analysis, the final response of every self-reporter in each of the questions was compared with the final response given by his proxy in the same question.

Results

For the analysis 352 sequences (i.e., 11 questions * 32 respondents) were taken into account.

Sequences types analysis

First, the sequence types produced by both types of informant were compared. Table 4 shows the results from this comparison.

Table 4. Percentage of different sequences produced by “proxies” and self-reporters

Sequence type	Self-reporters	Proxies reporter
Paradigmatic sequences	38	58
Non paradigmatic-non problematic	23	19
Non paradigmatic- problematic	39	23

$$\chi^2 = 15.706 \text{ p} < 0.001$$

As Table 4 shows, significant differences were found in the percentages of the types of sequence produced by both types of informant. The greater differences occur in the percentage of paradigmatic sequences, these being higher for proxy reporters. Also, self-reporters show a high percentage of non paradigmatic-problematic sequences. Thus, self-reporters not only deviate from

the paradigmatic pattern more often than proxy-reporters, but also these deviations are more often problematic.

Comparison between codes produced for proxies and self-reporters.

Next, the frequency of the answering behaviors was analyzed for both types of informant. First, the types of answer given by the respondent were observed. The answers were classified in two groups: problematic answers, where answers coded as mismatch answer, invalid answer, don't know answer and qualified answer were included; and non problematic answers composed by answers coded as adequate. The last one includes the sequences in which an adequate answer was given although other problematic answers occurred beforehand. Because of the existence of multiple behaviors in the same sequence, the total percentage can exceed 100 %. Table 5 shows the percentage of problematic and non problematic answers to each question of the questionnaire for both types of informant.

Table 5. Percentages of problematic and non problematic answers

Question number	Any problematic answer		Adequate answer	
	Self	Proxy	Self	Proxy
Q.1. "Habitual activities"	44	75	81	88
Q.2. "Speak"	44	19	69	81
Q.3. "Understand"	56	6	56	100
Q.4. "Use the phone"	69	44	81	75
Q.5. "Use the senses"	31	31	88	94
Q.6. "Learn"	38	25	94	100
Q.7. "Move the body"	31	19	94	100
Q.8. "Change posture"	44	19	63	94
Q.9. "Show affection"	31	13	88	94
Q.10. "Family relationships"	19	0	81	100
Q.11. "Sentimental relationships"	25	19	88	94

As Table 5 shows, in general, self-reporters show higher percentages of problematic answers than proxies except in question 5 "use the senses" where the percentage is equal for both (31%) and in question 1 "habitual activities" where proxies produce problematic answers in a very high percentage (75%). This last question and question 4 "use the phone" reached the highest percentages of problematic answers for both informants. Also, question 1 "habitual activities", question 3 "understand" and question 10 "family relationships" show the highest differences between both informants. In the first case, proxies gave more problematic answers than self-reporters, while in question 3 and question 10 self-reporters produced more problematic answers than proxies, who never produced a problematic answer in question 10. On the other hand, the percentage of non problematic answers, that is to say, of adequate answers was always greater for proxies except in question 4 "use the phone". The largest differences between both informants occurred in question 3 "understand" where proxies reached 100% of adequate answers.

In addition, the percentage of behaviors occurring while the interviewer was asking the question was analyzed for both types of informant. Table 6 shows the percentages of the occurrence of these additional behaviors.

Table 6. Percentages of the behaviors occurring while asking the question.

Question	Code			
	Request clarification		Interruption	
	Self	Proxy	Self	Proxy
Q.1. "Habitual activities"	6	6	6	0
Q.2. "Speak"	0	6	0	6
Q.3. "Understand"	6	0	0	6
Q.4. "Use the phone"	13	13	0	0
Q.5. "Use the senses"	6	0	0	0
Q.6. "Learn"	19	0	0	0
Q.7. "Move the body"	19	0	0	0
Q.8. "Change posture"	6	6	0	0
Q.9. "Show affection"	6	0	0	0
Q.10. "Family relationships"	0	13	0	0
Q.11. "Sentimental relationships"	13	0	0	0

As table 6 shows, the percentages obtained for the code 'request clarification' are in general higher for self-reporters than for proxies, except for the questions 2 "Speak" and 10 "Family relationships". A striking difference is that in the questions 6 "Learn" and 7 "Move the body" the percentages of occurrence are high for self-reporters whereas proxies never produce requests for clarification in these questions. As for the interruptions, they appear with higher frequency for proxies than for self-reporters, though the percentages are not high.

Convergence evaluation

After the informants' behavior analysis, the convergence between both types of informant was evaluated. Two approaches can be used to compute the

disagreement between both types of informant. The “traditional” approach calculates the percentage of disagreement taking all sequences into account. In doing that, the percentage of disagreement was 48% and the percentage of sequences with agreement in the answers given for both members of the same couple was 52 %.

The “traditional” approach uses all sequences no matter if the sequences are “problematic” or “non problematic”. When that approach is used, researchers miss that the convergence evaluation is not always possible or, at least, advisable. There are, for instance, situations in which one member of the couple did not give an answer or the answer given was not an adequate answer. For example, if in a yes/no question the proxy says “yes” and the self-reporter says “no” there is disagreement, but if proxy says “yes” and the self-reporter says “sometimes”, there is a mismatch answer, that is to say an answer which does not fit to any of the alternatives given. Traditionally, these situations have been considered as “disagreement situations”.

The behavior coding method allows us to optimize the convergence evaluation by only selecting cases in which both proxy and self-reporters gave an adequate answer. This selection is important for knowing the real percentage of disagreement and to filter the evaluation by removing situations in which a final adequate answer was not obtained. After removing these cases, the percentages of disagreement were reduced to 19%.

Furthermore, the “disagreement in answers” sequences found when either proxy or self-reporters give a non adequate answer, can be categorized into three groups: a) “Self non adequate”, when the self reporter did not give an adequate answer but the proxy did; b) “Proxy non adequate”, when the contrary occurred; and c) “Both non adequate”, when neither of them gave a final adequate answer. Table 7 shows the percentages for each of these categories together with the percentages of agreement and disagreement when both types of informant ended up giving an adequate answer.

Table 7. Percentage of sequence with agreement and disagreement

	Agreement/disagreement	Percentage
Adequate answers	Agreement in answers	52
	Disagreement in answers	19
Non adequate answers	Self non adequate	23
	Proxy non adequate	6
	Both non adequate	0

Table 7 shows how the percentage of disagreement in the answers descends to 19% due to the percentage of non adequate answers (29% in total) mainly on the part of the self-reporter (23%). To carry out the convergence analyses between the answers provided by self-reporters and proxies, only the percentages of adequate answers were included.

In addition, the percentages of agreement and disagreement throughout the set of target questions were calculated. Table 8 presents the percentages of agreement/disagreement in answers calculated by only counting final “adequate” answers. Table 8 also shows the percentages of non adequate answers in order to detect cases with high differences between both types of informant.

Table 8. Agreement between proxy and self-reporter answers

Agreement/ Disagreement	Questions numbers										
	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10	Q.11
Agreement in answers	38	38	31	44	63	75	56	38	56	56	75
Disagreement in answers	19	13	19	13	19	19	31	19	25	25	6
Self non adequate	25	38	50	25	13	6	13	38	13	19	19
Proxy non adequate	19	13	0	19	6	0	0	6	6	0	0

Table 8 shows differences across the different questions. A high percentage of disagreement exists for question 7 “Family relationships” (31 %), while questions 6 “Learn” and 11 “sentimental relationships” achieve the highest degree of agreement (75 %). In relation to non adequate answers, the largest percentages are observed in self-reporters for all the questions. For example, self-reporters have the highest percentage of non adequate answers (50%), while the percentage achieved by the proxies in the same question was 0.

Discussion

The objectives of the study were to analyze the behavior of proxy reporters and to evaluate the convergence between the answers given by self-reporters and proxies to a disability questionnaire by means of a behavior coding method. The rationale behind applying a behavior coding method was to improve the evaluation of convergence by analyzing only final adequate answers i.e. answers that are directly interpretable as an answer, given by both types of informant.

The analysis of the behavior of both types of informant showed that the percentage of paradigmatic sequences in both groups can be considered

adequate. Van der Zouwen and Smit (2004) found in a study 29.8 % of paradigmatic sequences. However, this percentage was higher in the group of proxy reporters (37.5% versus 57.95%). The opposite difference was found with regard to the percentages of non paradigmatic-problematic sequence (39.20% versus 23.30%). The percentage of non paradigmatic-problematic sequences is slightly higher than the percentages found in the bibliography. Dijkstra and Ongena (2006) found percentages between 22 and 37.9 % of problematic sequences in the analysis of five different surveys.

The extent to which both types of informant perform the stages of the cognitive “question-and-answer” process may explain the differences in the percentages for the types of sequence. According to the Krosnick’s theory of “optimizing vs. satisficing”, the involvement of the respondents when answering survey questions depends on three points: task difficulty, respondents’ ability and respondents’ motivation (Krosnick, 1999). Based on the findings of behavior coding in this study, the difficulty of the task and the motivation for responding to questions could be different for both types of informant. Self-reporters may have more information available in the episodic memory, which, in contrast to first impressions, could make it more difficult to translate to a response category. Among the aspects which may influence the participants’ motivation, the personal importance of the question’s topic to the respondents could be the most important. In this study, the questionnaire topic might be more important for the self-reporter because they report on their own situation. On the other hand, proxies report on the situation of another person. All these aspect taken together could increase the chance of a high number of problematic answers for self-reporters, increasing the probability of non paradigmatic sequences.

In the specific analysis of the appearance of problematic and non problematic answers, self-reporters showed a higher percentage of problematic answers than proxies. In this analysis, it is important to point out three main results. Firstly, questions with a high percentage of problematic answers for

both informants, such as question 4 “use the phone”, might indicate possible problems with the wording or the format of the question. Question 4 has a complex format because it includes instructions about some information that the respondent should exclude while answering, which is a challenging cognitive task. Secondly, in questions in which proxies have obtained higher percentages of problematic answers, such as question 1 about limitations to habitual activities, the percentage of problematic answers might be due to the proxies’ lack of detailed information about the topic. Finally, there are questions in which self-reporter obtained higher percentages of problematic answers, such as question 3 and 10.

These results might be explained by the optimizing theory. According to the optimizing theory a respondent who is optimizing would carefully assess the appropriateness of each response before selecting one. In contrast, a respondent who is satisficing could simply choose the first reasonable response (Krosnick, 1999). Considering that self-reporters are optimizing, because the topic is more important for them and they should be more motivated, it is possible they analysed the alternatives, which are “yes” and “no” (all questions are yes/no questions) determining that neither of them were completely adequate. In this case, the self-reporter probably gave an invalid or mismatch answer. On the other hand, proxies are satisficing and they maybe selecting the closest alternative to the real situation, and for this reason they achieved 100% of adequate answers, but which may not necessarily be the most correct answer.

The convergence analysis showed that the percentage of agreement was 52% and how the percentage of disagreement falls from 48% to 19% when counting only final adequate answers. Both results were obtained by analyzing “comparable” answers given by proxies and self-reporters. On the other hand, the convergence analysis for each of the target questions showed higher percentages of agreement in questions 6 and 11, whereas question 7 reaches the highest percentage of disagreement. The content of question 6 was about difficulties when writing, copying, counting or using everyday utensils. As

Magaziner, Speaar, Hebel and Gruber-Baldini (1996) point out, proxies inform accurately about physical or daily tasks, and this might be the reason for the high agreement found. Question 11 is about sentimental and sexual relations. In this case, the intimate content of the question could cause both informants to become satisficers, influenced by social desirability, thereby reaching a high percentage of agreement. In relation to disagreement situations, question 7 obtained the highest percentage. This question had a complex format and its content was also complex. It could be possible both informants went through a different satisficing process (i.e. strong or weak satisficing) to answer this question. As Schwarz and Wellens have already pointed out, the information retrieval and judgment process is likely to proceed differently for proxy versus self-reporters.

In conclusion, in spite of the results found by Loftus, Smith, Klinger and Fiedler (1992) and Todorov and Kirchner (2000) in their studies with proxies and self-reporters, the results show a better behavior of the proxy reporters from an interviewer-respondent interaction perspective. However, although the answers given by proxies have been adequate in a high percentage and the behavior while asking the question has been less problematic, we have to take into account that satisficing can not be detected as clearly by means of behavior coding. Thus, it is possible that although they gave adequate answers, the proxies were satisficing. For this reason, although it is possible to say the convergence between proxies and self-reporters exists because a high percentage of agreement has been obtained, it is necessary to have more detailed information on the cognitive process taken by both informants to assure that their answers are really equally valid. Thus future research could be focused on applying other procedures, such as cognitive interviews, to obtain more information.

In relation to the usefulness of behavior coding, this procedure has made two fundamental contributions. First, it enables analyses with the information obtained directly from the interaction, i.e., what has actually happened during

the interview. Traditionally, this type of analysis has been carried out using the information registered by the interviewer in the questionnaire, which could be more biased than the information proceeding directly from the record of the interaction. Second, behavior coding has allowed selecting the analysis only for those cases in which both informants had given an adequate answer. Including only these cases yields a truly fair comparison of proxies and self-reporters, since the interactional situation of the answer was taken into account.

References

- DeMaio, T.J., Rothgeb, J., & Hess, J. (1998). Improving survey quality through pretesting. Washington, DC: U.S. Bureau of the Census. Retrieved July 30, 2009, from http://www.amstat.org/Sections/Srms/Proceedings/papers/1998_007.pdf
- Duncan, P., Min Lai, S., Tyler, D., Perera, S., Reker, D.M., & Studenski, S. (2002). Evaluation of Proxy Responses to the Stroke Impact Scale. *Stroke*, 33, 2593-2599.
- Dijkstra, W. (2008). Sequence Viewer (version 4.4a). Free University of Amsterdam, Netherlands.
- Dijkstra, W., & Ongena, Y. (2006). Question-answer sequences in survey-interviews. *Quality & Quantity*, 40 (6), 983-1011.
- Forsyth, B., Levin, K., & Fisher, S. (1999). *Test of an appraisal method for establishment survey questionnaires*. Proceeding of the ASA Section on Survey Research Methods. Alexandria, VA: American Statistical Association.
- Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology* 50, 537-567.
- Loftus, E.F., Smith, K.D., Klinger, M.R., & Fiedler, J. (1992). Memory and mismemory for health events. In: J.M. Tanur (Ed.) *Questions About Questions: Inquiries Into the Cognitive Bases of Surveys* (pp. 102-137). New York, NY: Russell Sage.
- Magaziner, J., Speaar, S., Hebel, J.R. & Gruber-Baldini, A. (1996). Use of "proxies" to measure health and functional status in epidemiologic studies of community-dwelling women. *American Journal of Epidemiology*, 143 (3), 283-292.
- Oksenberg, L., Cannell, C. & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics* 7 (3), 349-365.

- Ongena, Y.P. (2005). Interviewer and Respondent Interaction in Survey Interviews. Unpublished doctoral dissertation. Amsterdam Vrije Universiteit.
- Ongena, Y.P. & W. Dijkstra (2006). Methods of Behavior Coding of Survey Interviews. *Journal of Official Statistics* 22, 419-451.
- Pickard, A.S., Johnson, J.A., Feeny, D.H., Ashfaq, M.D., Carriere, K.C. & Abdul M.N. (2004) Agreement Between Patient and Proxy Assessments of Health-Related Quality of Life After Stroke Using the EQ-5D and Health Utilities Index. *Stroke*, 35, 607-612.
- Schaeffer, N.C., & Maynard, D.W. (1996). From Paradigm to Prototype and Back again: Interactive Aspects of Cognitive Processing in Standardized Survey Interviews. In N. Schwarz and S. Sudman (Eds.), *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp.367-391). San Francisco: Jossey-Bass.
- Snijkers, G. (2002). *Cognitive Laboratory Experience: On Pre-testing Computerised Questionnaires and Data Quality*. Ph.D thesis. Utrecht University, Utrecht, and Statistics. Netherlands, Heerlen.
- Sunghee, L. Mathiowetz, N.A., & Tourangeau, R. (2007) Measuring disabilities in surveys: Consistency over the time and across respondents. *Journal of Official Statistics*, 23 (2), 163-184.
- Schwarz, N, & T. Wellens (1997) Cognitive Dynamics of Proxy Responding: The Diverging Perspectives of Acotrs and Observers. . *Journal of Official Statistics*, 13 (2), 159-174.
- Tafforeau, J., Lopez, M., Tolonen, A., Scheidt-Nave, C. & Tinto, A. (2006). Guidelines for the development and criteria for the adoption of Health Survey instruments.

- Todorov, A., & Kirchner, C. (2000). Bias in “proxies” Reports of Disability: Data From the National Health Interview Survey on Disability. *American Journal of Public Health, 90* (8), 1248-1253.
- Tourangeau, R., Rips, L.J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.
- Tourangeau, R. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research, 15*, 3-7.
- Van der Zouwen, J. & Smit, J. H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behaviour Codes and transcripts of Question-Answer Sequences: A Diagnostic Approach. In S. Presser, J. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin and E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp.109-130). New York: Wiley.
- World Health Organization (2009). ICDH-2: International Classification of Functioning, Disability and Health (ICF). Geneva, Switzerland: Author.

CHAPTER 3:

Obtaining validity evidence by cognitive interviewing to interpret psychometric results

Padilla, J.L., Benítez, I., y Castillo, M. (2012, under review). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Special Issue Survey Research Methods*.

Abstract

The latest edition of the *Standards for Educational and Psychological Testing* (1999) suggests resorting to the empirical and theoretical analysis of respondents' response processes in order to obtain evidence about the fit between the intended construct and the response process actually produced. The main aim of this paper was twofold. First, the study was intended to demonstrate how cognitive interviewing can be used to gather such validity evidence, and secondly, to analyse the usefulness of the evidence provided by cognitive interviews for interpreting the results from traditional psychometric analysis. The usefulness of the Cognitive Interviewing Reporting Framework (CIRF) for reporting the cognitive interviewing findings was also evaluated. As an empirical example, we tested the (Spanish language) APGAR family function scale consisting of 5 items in a rating scale format with 3 response options. A total of 21 pretest cognitive interviews were performed, and then psychometric analyses were conducted of data from 28,371 respondents who were administered the APGAR scale within a national health survey. Cognitive interviewing evidence was used to interpret differences in psychometrics for multiple respondent groups. We conclude by presenting pros and cons of the CIRF as a quality framework.

Keywords: Cognitive interviewing, response processes, validity, psychometrics.

Introduction

The use of cognitive interviewing as a method for evaluating the quality of questions included in surveys has become more and more popular (Castillo, et al., 2010). As Conrad, Blair and Tracy (1999) indicate, cognitive interviewing is the pre-test questionnaire method which is the meeting point between cognitive psychology and survey methodology. In the latter area, the main goal of cognitive interviewing is to identify potential sources of measurement error and improve the wording of survey questions (Tourangeau, 2003). Cognitive interviewing is a powerful method for understanding the thought processes made by the respondents when answering survey questions (Beatty, & Willis, 2007).

Although the application of cognitive interviewing has been mainly related to questionnaires included in surveys, it can also be implemented for the assessment of psychological scales. Conrad and Blair (2009) pointed out a number of conditions that enable cognitive interviewing in obtaining evidence about the non-automatic processing of survey items. Summing up, assuming the "non-automatic nature" of the thought processes implies that respondents are able to discuss some of the thought processes that have taken place in their short-term memories.

Despite this potential, application of cognitive interviewing for the assessment of psychological scales has been rare. Poole, Murphy and Nurmikko (2009) evaluated the NePiQoL scale of the "quality of life" construct through cognitive interviewing to identify items with redundant content and obtain evidence of content validity. Further, Knafl et al. (2007) used the cognitive interviewing method to develop and analyze a parental management psychological scale of childhood chronic conditions.

Cognitive interview data have also been used to facilitate the interpretation of quantitative data obtained by psychometric analysis. DeWalt, et al. (2007) used cognitive interviews to evaluate the ambiguity, understanding and relevance of a set of psychological items. Hardinson and Neimeyer (2007)

evaluated the convergence of quantitative and qualitative information provided by different methods used to analyse the process that people carry out when building individual mental models. In that study, the authors noted that using a single method, whatever it may be, offers an overview of the interviewee's point of view and also reduces the complexity of the task. Moreover, in the study carried out by Olt, Jirwe, Gustavsson and Emami (2010), different procedures were used to obtain information on the quality of a cultural competence scale that had been translated into Swiss. The authors used quantitative methods including psychometric statistics such as the Cronbach's alpha coefficient to estimate reliability and confirmatory factor analysis, together with qualitative methods such as cognitive interviews. In this case, the authors demonstrated the importance of the compatibility of the methods through the richness of the obtained results. In a similar approach, Sarkisian, et al., (2002) applied cognitive interviews, focus groups, and multi-trait scaling analysis in the evaluation of ERA-38 scale (Expectations Regarding Ageing), finding convergence in terms of evidence of construct validity.

The Cognitive Interviewing Reporting Framework (CIRF)

Boeije and Willis (this volume) propose the CIRF as a quality framework that cognitive interviewers can use when reporting cognitive interviewing studies. Underlying the CIRF is the assumption that cognitive interviewing is in essence a qualitative method. Thus, cognitive interviewers can benefit from the learned lessons in the qualitative research field to improve their research projects by using the CIRF as a guide. Taking the unique characteristics of the cognitive interviewing into account, the CIRF authors included 10 major categories items in the CIRF.

Overall objectives

As it was stated before, the evaluation of psychological scales has been performed almost exclusively using psychometric analysis, undervaluing the input of more qualitative approaches. Hence, the main objective of this study is to determine whether the evidence provided by cognitive interviews facilitates

the interpretation of the results obtained from psychometric statistics. To this end, two studies have been carried out using the responses of two groups of participants to the APGAR family function scale. This scale was developed by G. Smilkestein in 1978 to assess the construct of family support in health surveys by evaluating the perception respondents have of the support they receive from family members. Study 1 investigates the use of traditional psychometric analyses by comparing two groups of people: residents in single person, and those within multi-person households. The logic behind this division is that these groups can differently interpret relevant aspects of the construct "family function" as it was defined by Smilkestein (1977).

In Study 2, we present evidence based on the response processes obtained through cognitive interviewing. The analysis of these response processes allows us to investigate potentially different interpretations of the construct, and the different relevance of the indicators reflected in the scale items. Then, we related the evidence provided by cognitive interviewing with the various psychometric test results for both groups of respondents. Finally, we endeavoured to use the CIRF to report cognitive interviewing results when the aim of the cognitive interviewing study is not to test survey questions, but to obtain validity evidence and insights for helping researchers interpret psychometric results. Hence, Study 2 was written following the ten numbered CIRF categories and recommendations.

Study 1: The analysis of the psychometric characteristics of the items included in the APGAR scale.

The aim of study 1 was to analyse the psychometric characteristics of the APGAR scale items (Smilkestein, 1978). The psychometric analysis was carried out by separating the total sample according to the variable "type of home": single person or multi-person household.

Method

Participants

26413 Spanish people responded to the APGAR scale included in the adult questionnaire of the Spanish Health Survey (Spanish Ministry of Health and Consumption, 2006). Table 1 shows the main demographics with regard to the variables gender, marital status, and type of home.

Table 1. Demographic characteristics of participants

Variables	Frecuency	Percentege
Gender		
- Male	10298	39.0
- Female	16115	61.0
Marital status		
- Single	6380	24.2
- Married	15215	57.6
- Widowed	3445	13.0
- Legally separated	772	2.9
- Divorced	601	2.3
Type of household		
- Uni-personal	10042	38.0
- Multi-personal	16371	62.0

As Table 1 shows, 61% of the participants were women and 39% men. With regard to marital status, the majority of the participants were married at the time of the survey (57.6%), while divorced was the lowest percentage (2.3%). Finally, 38% of the participants lived in a one-person home and 68% shared their household with other people.

Materials

Psychometric analysis was conducted on respondents' answers to the Spanish version of the APGAR family support scale (Bellón, et al., 1996)

included in the Spanish Health Survey. The APGAR scale was designed to assess the perception of family members on family support by examining their satisfaction with family relationships. The APGAR scale is used in clinical practice by family physicians as a tool for quickly and easily gathering information on the family situation and its possible role in the origin and resolution of conflicts (Bellon et al., 1996). The scale has five components of family support from five Likert items: adaptability, partnership, growth, affection and resolve (Smilkstein, 1978). Table 2 presents the APGAR questionnaire items in its original version.

Table 2. Original APGAR items

	Almost always	Some of the time	Hardly ever
1. Are you satisfied with the help you receive from your family when you have a problem?	<input type="checkbox"/> 0	<input type="checkbox"/> 1	<input type="checkbox"/> 2
2. Are you satisfied with the time you and your family spend together?	<input type="checkbox"/> 0	<input type="checkbox"/> 1	<input type="checkbox"/> 2
3. Do you feel your family loves you?	<input type="checkbox"/> 0	<input type="checkbox"/> 1	<input type="checkbox"/> 2
4. Do you talk together about problems you have in home?	<input type="checkbox"/> 0	<input type="checkbox"/> 1	<input type="checkbox"/> 2
5. Important decisions are made by all of you together in home?	<input type="checkbox"/> 0	<input type="checkbox"/> 1	<input type="checkbox"/> 2

The Spanish version of the APGAR scale has shown high internal consistency values (Cronbach's alpha coefficient of 0.84), revealing a one-dimensional structure in the factor analysis (Bellon et al., 1996). The brevity and usefulness of the APGAR scale at various levels have made it a frequently used instrument (Wolnitzki et al. 1989; Austin, & Huberty, 1989; Mauricio Romero, et al., 2001; Montecinos, 2007).

Data collection

Data from the responses to the APGAR scale are available on the database: (<http://www.msps.es/estadEstudios/estadisticas/solicitud.htm>). The database was reviewed to include those participants who had responded to all items of the scale in the analysis.

Data analysis

First, a descriptive analysis of the distribution of total scores was performed together with an analysis of the individual APGAR items. Then, the reliability and dimensionality of the responses were analysed. All tests were conducted by dividing participants according to type of household. Analyses were performed with the *Statistical Package for Social Sciences* (SPSS v.16).

Results

The mean and standard deviation values of the distributions of the responses to the items are shown separately in Table 3 for each of the groups.

Table 3. Descriptive statistics for the APGAR items.

Item	Single person household			Multi person household		
	Mean	SD	DI	Mean	ST	DI
1	1.86	0.429	0.675	1.90	0.344	0.500
2	1.76	0.558	0.726	1.89	0.359	0.554
3	1.70	0.626	0.655	1.87	0.401	0.485
4	1.70	0.582	0.562	1.72	0.548	0.304
5	1.91	0.344	0.655	1.97	0.205	0.515

Note: SD = Standard Deviation; and DI = Discrimination Index

The mean responses to all items were lower in the group of respondents who lived alone. Items 3 and 4 obtained the lowest mean values in this group of participants. All the differences between the mean values of both groups were

statistically significant for a confidence level of 99%. On the other hand, the standard deviation values were lower in the group of respondents living in multi-person households in all cases, indicating a lower variability in the responses of this group. The discrimination indices of the items were within the typical values for these scales, with higher values in the group of participants with single person households. Again, in item 4 lower discrimination index values are recorded in both groups. The Cronbach's alpha coefficient was 0.83 for the group of single person households and 0.68 for multi-person households. By calculating the Feltd's *W* statistic (Feltd, 1969), it was found that the difference between the alpha coefficient values in both groups was significant ($W = 1.94, p < 0.001$). This difference may indicate that the answers to the items of respondents in single person households are significantly more homogeneous than those in the multi-person households.

The dimensionality of responses to items in both groups was analysed using exploratory factor analysis of principal axes. The values of the Bartlett test of sphericity (single person: $\chi^2_{(10)} = 21,797.48, p < 0.0001$; and multi-person: $\chi^2_{(10)} = 16,845.86, p < 0.0001$), together with the values of the KMO index (single person: .81; and multi-person: .80), guaranteed that both the two matrices of correlations between items were factorable. Table 4 shows the eigenvalues and the percentage of explained variance for each factor.

Table 4. Exploratory factor analysis results.

Groups	Factor 1		Factor 2		Factor 3	
	Eigenvalue	Explained variance	Eigenvalue	Explained variance	Eigenvalue	Explained variance
Single person household	3.127	62.549	0.692	13.830	0.573	11.466
Multi person household	2.449	48.975	0.855	17.100	0.721	14.416

As shown in Table 4, in both groups a single factor was obtained with eigenvalues greater than one that accounted for 62.54% of the variance in the case of participants with single person households, and 48.97% for the multi-person. Both values confirm the unidimensionality of the scale according to the criteria established in the literature (Carmines, & Zeller, 1979; Reckase, 1979). Table 5 shows the commonalities and the factor loadings of the items for each of the groups.

Table 5. Factor loadings and communalities of the APGAR items

Items	Factor loadings		Communalities	
	Single person household	Multi person household	Single person household	Multi person household
1	0.796	0.673	0.730	0.533
2	0.815	0.734	0.760	0.592
3	0.737	0.645	0.657	0.499
4	0.614	0.352	0.409	0.144
5	0.754	0.641	0.655	0.454

The communalities of the items after extraction, i.e. the variance in the answers that each item shared with others, were higher in the group of single person households for all items. Moreover, the factor loadings of the items in the first factor were also higher in all cases in the group of participants with single person households. Again, the most notable difference appears in item 4, for which a very low load factor was obtained in the multi-person group, while in the single person household group, despite reaching the lowest value it was more similar to the other items. Both results show that the unidimensionality is clearer in the case of participants living in single person households, and that item 4 could represent another dimension, especially in the case of the multi-person household group.

Conclusions

The results of psychometric analysis performed show that the performance of the AGPAR scale items is different in the two groups of respondents. Such differences were detected in both the mean values and standard deviations of responses to the items, as well as in the levels of item discriminations. The results indicate that participants with multi-person households scored higher on the scale of "family support" than participants with single person households, although in the latter group the items discriminated better, possibly because there was greater variability in the responses of these participants. On the other hand, similarities were also found between the subgroups, because in both cases item 4 obtained the lowest average and discrimination index. This item was also the one which had the lowest factor loading in factor 1 and the lowest in the commonalities, especially in the multi-person homes group.

With regard to the consistency of responses and their dimensionality, the results confirmed the unidimensionality of the scale. Nevertheless, the significant differences found between the Cronbach's alpha coefficient values and the explained variance percentages suggest that the responses of the group of participants in single-person households are more homogeneous.

Study 2: Use of cognitive interviews to gather evidence about the answer process carried out by the respondents

To reiterate, the following write-up of Study 2 was done using the ten major CIRF categories - although the ordering of these varied from that used within the CIRF system.

Research objectives

The aim of study 2 was to gather validity evidence by means of cognitive interviews for interpreting the results of the psychometric analysis performed. The possible differences in the answer processes of the participants from single

person and multi-person households could reveal different interpretations of the "family function" construct, and the understanding of the indicators reflected in the items.

Method

Participant selection.

The cognitive interviews were conducted with 10 men and 11 women, aged between 20 and 67 years (seven participants, aged between 20 and 35 years old, seven participants between 36 and 50, and seven participants between 51 and 67 years old). Of these, 12 were single, six married and three divorced. As for the type of housing, nine people with one-person households and 12 multi-person households participated. Participants with different marital status and who were living in different type of household were recruited trying to capture different interpretations of the "family function" construct.

Materials

Interviewers used an interview protocol which included follow-up probes. The follow-up probes included in the protocol were applied retrospectively, i.e. first the complete APGAR scale was administered, and then the follow-up probes. The retrospective application of the probes is appropriate when the presentation of the items is desired to be as realistic as possible (Willis, 2005). In this case, the retrospective application enabled the administration of the APGAR scale carried out in Studies 1 and 2 to be comparable.

Research Design

To carry out the cognitive interviews a "probing based" paradigm was applied (Beatty & Willis, 2007) which included general and specific probes. This paradigm is characterized by the fact that the questions and follow-up probes guide the interaction, giving the interviewer the freedom to explore relevant issues.

Ethics and data collection

First, participants were informed about the purpose of the study. To motivate participants, they were told how important the interviews will be to improve a national health survey promoted by an official body. The interviewers told the participants the information provided by the survey will be used by policy makers. In addition, each participant was rewarded with 30 Euros. The interviews were conducted individually by four trained and experienced interviewers (three females and one male). The interviews were recorded on audio and video with the consent of the participants and took place in a cognitive laboratory. The participants were guaranteed confidentiality and that the data would be solely used for purposes related to research.

Data analysis

The analysis of the cognitive interview data was conducted following the approach, in several stages, developed by Miller (2007). In the first stage, the interviews were analysed individually in order to reveal the participants' interpretations of the items (in this case taking all the items as a block, as the APGAR is a multi-item scale). From this data set three main themes were established, within which the participants developed different subthemes. During the second stage, the interpretations made by different groups of participants defined by the type of household were compared. This comparison made it possible to test whether the problems with the scale items were specific to a group or common to all participants. Differences between groups were analysed based on the interpretations of the indicators developed in the replies of the participants. Two independent analysts analysed the results, and these analyses were subsequently compared and discussed at a meeting to achieve maximum consensus.

Findings

The evidence obtained from the cognitive interviews is presented in two parts. The first part shows the themes developed by the participants during the whole of the cognitive interview. The second part presents the differences

detected between the interpretations made by participants in terms of household types: single-person and multi-person. Throughout the study participants are identified by codes made up of different acronyms that correspond to their characteristics: 1) number of interviews (1-9 = participants with single person households, 10-21 = participants with multi-person households), 2) gender (M = Man, W = Woman), 3) age, 4) marital status (S = Single M = Married; D = Divorced), and 5) Type of household (U = single person, M = multi-person)

First stage: Themes developed by the participants.

The follow-up probes included in the interview protocol were designed to obtain evidence on three general themes: the “concept of family”, “asking for help”, and “making decisions”. These three general themes were included on the advice of the experts responsible for the survey and after a review of the literature on the construct "family function." During the first stage of analysis, the subthemes obtained from the comments of the interviewees were grouped within each general theme.

First, in relation to the concept of "family", the participants talked about the number and the members of their family, the time spent with family and the frequency of contact with those members. As for family members, six of the participants felt that family members are the people who they live with, although only one of the participants made it explicit that they had thought only of these people when responding to the items of the scale (14.W.38.D.M). Nevertheless, the remaining 15 participants spoke of other family members with whom they did not live.

Further, when discussing “asking for help” from family members, participants commented on two aspects. On the one hand, they discussed the different types of support they receive from their family, and on the other, they listed the members who provided such assistance. Regarding the type of aid: four participants spoke in general without making any specification,

commenting for example that in difficult situations family members have always helped (5.W.45.D.U; 8.M.36.S.U; 10.M.35.S.M and 16.W.40.S.M), two participants spoke only about economic issues, 10 spoke only of moral or psychological support (such as giving advice, being listened to, etc.) and five spoke both of economic and psychological support. Two of the participants also discussed the importance of feeling that others care for you, for your health and well-being (3.M.64.S.U and 14.W.38.D.M.). Regarding the people providing aid, most of the participants mentioned their parents, siblings, children and partners. Only one person commented that he received no help from anyone and had lived difficult situations and felt alone (13.W.67.M.M) and two participants referred to people outside the family such as friends and neighbours (3.M.64.S.U and 9.M.30.S.U).

Finally, in terms of "family decision making", participants talked about both the type of decisions that are often discussed in the family and the members who are part of those decisions. On the types of decisions, statements of the interviewees can be divided into five blocks: holidays and travel, purchase of household services such as internet, daily purchases or occasional (e.g. a car or a home), economic affairs such as loan applications, and important changes such as change of residence, work or children's school. In relation to members who take decisions, nine of the participants made reference to decisions outside the home, that is still taken with people who no longer live together but lived together in the past. Also, one participant said that before making a decision he would consult with people outside the family to get their opinion (3.M.64.S.U).

Second Stage: Comparison of the interpretations made by the participants by type of household.

Following the themes, an analysis was conducted to detect differences in the interpretations made by participants by household types: one-person and multi-person. Firstly, differences were observed in members who were

included in the family. Of the nine participants with single person households, seven responded thinking about their parents, children or siblings, while one of the participants listed their friends when asked about members of his family and another said the following about his family "currently none as I live alone (3.M.64.S.U). In none of the cases the participants were living with these family members, because they lived in single person households. As for participants in multi-person households, only five of the 12 responded to the scale thinking only of the people who they lived with, while the seven remaining mentioned members of the family outside the home. Example 1 shows some of the statements of the participants.

Example 1. Family Concept of participants

"There are seven, and four nephews ... eleven, and counting the husbands and wives, there are then fourteen. We see each other very often, and whenever I want to see them, I see them all. I'm always in touch with them a lot" (8.M.36.S.U)

"There are five brothers and between nephews, brothers ... 12. With my son I have telephone contact, but mostly I have contact with my mother and my sister, because we live in the same area." (10.M.35.S.M.).

Regarding the request for assistance, differences between participants in both groups were found. Table 6 shows the sub themes developed by each participant. Participants with multi-person households are marked in italics.

Table 6. Themes developed concerning providing help

From whom...	Kind of help			
	Financial	Psychological	Both financial and psychological	Worry about him / her
Parents	2.M.36.S.U.		18.W.30.S.M.	
Siblings	6.M.32.S.U.	14.W.38.D.M.		14.W.38.D.M.
Parents and siblings		3.M.64.S.U. 11.W.28.D.M. 19.M.20.S.M.	4.M.40.S.U.	3.M.64.S.U.
Partner		12.W.46.M.M. 20.M.65.M.M.		
Children		7.W.67.S.U. 17.W.53.M.M.		
No family members		3.M.64.S.U.	9.M.30.S.U.	3.M.64.S.U.

As shown in Table 6 regarding the type of aid provided, it was observed that single-person household participants identified a greater variety of types of support than those from multi-person households. Thus, participants in single-person households made reference only to financial or psychological assistance or both at once, while the multi-person households at best mentioned only psychological assistance (8 of 10 participants), or in two cases combined with financial aid. In addition, participants in single-person households were those who also included reference to interest in their situation, to people worrying for them, among the types of assistance. Moreover, in relation to members from whom they sought help, it was found that participants with single person households never referred to couples and were also the only ones who

mentioned people outside the family. Example 2 shows a situation described by one member of each of the subgroups.

Example 2. Type of aid provided by members of each group.

"I try to do things for myself, I try not to have to ask for help, but the times I had to ask have been for economic reasons, for example when we were forming the company." (2.M.36.S.U.).

"They give you advice when you have a problem, help you, support you, they are with you, they don't leave you alone, they are with you when you need it." (12.W.46.M.M).

This data could indicate that for the group of participants from single person households the interpretation of the concept of "aid" in terms of the kind of help sought and from whom is more varied and heterogeneous than in the case of participants from multi-person households where the focus is more centred on the type of psychological assistance and the household members. In addition, participants from single person households more often understood the concept of aid within the meaning of material need, compared to the emotional need (as in the case of participants from multi-person households), indicating that the construct indicators could be different for both groups.

There were also differences in the responses of participants about the "decision-making" subthemes. Table 7 shows the topics discussed by each interviewee grouped by type of decision and the family members with whom they took them.

Table 7. Themes developed concerning making decisions.

Kind of decision	Members who make decisions			
	Interviewee	Partner and child	Household members	Former homes
Holidays	1.W.48.S.U.		16.W.40.S.M.	6.M.32.S.U.
	9.M.30.S.U.		19.M.20.S.M.	20.M.65.M.M.
Household services		21.M.51.M.M.		6.M.32.S.U.
Purchases	1.W.48.S.U.	21.M.51.M.M.	16.W.40.S.M.	4.M.40.S.U.
	9.M.30.S.U.		18.W.30.S.M.	6.M.32.S.U.
				13.W.67.M.M.
Economic affairs	1.W.48.S.U.	21.M.51.M.M.		
Important changes				4.M.40.S.U.
Missing				2.M.36.S.U.
				5.W.45.D.U.
	3.M.64.S.U.	12.W.46.M.M.		7.W.67.S.U.
		15.W.51.M.M.		8.M.36.S.U.
				10.M.35.S.M.
				17.W.53.M.M.

As shown in Table 7, six of the nine participants from single member households related how decisions were made in their former homes, i.e., situations that occurred with people who they had lived with in the past but were not living with today. The three remaining single-person participants reported that they made decisions for themselves, one of them specifying consulting occasionally with his friends and acquaintances. With regard to participants from multi-person households, only four of the twelve made reference to the decisions of their former homes (compared to two thirds of

participants from single person households), while seven of them spoke of the decisions taken together with the people they live with. Example 3 shows some of the comments made by the interviewees in relation to decision taking in the home.

Example 3. Statements on the decision-making.

"I suppose my parents make the decisions together. I guess my brother is also involved now, I'm further out, but there are cases in which decisions are made together" (4.M.40.S.U).

"Well, my daughter wants to buy a house and asks my opinion and that I go with her. With my husband, the children's schedule, many things that I consult with him about, we speak before doing anything. If we have to buy anything, or the weekend we want to go somewhere. The decisions are usually made by my oldest son, my husband, me and our daughter, the other girl is very small and is fine with everything" (12.W.46.M.M).

This evidence points to a greater heterogeneity with respect to the people who make decisions in the group of participants of single person households, compared to the focus on household members from multi-person households.

Discussion and Conclusions

The results gathered provided information about the interpretations made by participants from both groups of the indicators of the "family function" construct. These differences focused on the aspects considered when discussing the types of assistance and the decision-making processes. Participants from single person households showed a greater variety in their conversation for both aspects. Thus, participants from single person households related how they take their own decisions together with the decision-making processes that they performed in their former homes, that is, when they lived with their parents before becoming independent or when their children were still living at home. However, participants from multi-person households spoke mostly

about experiences in their current home. On the other hand, it was observed that participants from single person households gave more weight to "material" issues along with those of an "emotional" character, while those from multi-person households focused on the "psychological help".

This data shows how participants from different types of households have used different frameworks to respond to the scale, and additionally they have given different weight to the aspects considered when evaluating family support with regard to the assistance they receive from their members.

The main objective of this study was to show how the evidence obtained through cognitive interviews helps to interpret the results of psychometric tests. On the one hand, cognitive interviewing has yielded evidence of validity about the response processes of respondents to the items on the APGAR scale (Smilkstein, 1978). Evidence obtained through cognitive interviewing allows us to understand the interpretations of both the construct "family function" and the indicators of this construct reflected in the items of the scale. On the other hand, evidence of validity obtained have allowed the interpretation of the differences between the results of traditional psychometric tests comparing two groups of respondents whose situations could be associated with differential interpretations of the "family function" construct.

The different interpretations identified by cognitive interviewing have matched differences in the values of the psychometric results. For example, the greater variety in the interpretation of the types of assistance and the people who share or shared decision making, is associated with greater heterogeneity in the responses of the group of participants from single person households and, as a result, with higher values of discrimination indices of the items, internal consistency of the scale, and unidimensionality.

With regard to the APGAR scale, it is also important to note the caution that must be taken if interpretations that include participants with different types of households are desired, i.e., cognitive interviews have provided information on

situations in which the scale is not appropriate because the interpretation of respondents in these specific circumstances vary, while they should remain stable across different groups.

Cognitive interviewing can provide useful information about the respondents' response processes when responding to the items of a psychological scale, provided that these response processes are not automatic, and that participants use information stored in the short term memory (Conrad and Blair, 2009). Cognitive interview data have revealed their ability to provide useful validity evidence of the response processes for the interpretation of the measurements obtained by psychological scales. However, they also have some disadvantages, among which the most important are the lack of objectivity and consistency (Conrad, Blair, and Tracy, 1999). These two problems are related to the absence of a theoretical basis to guide the analysis of data from cognitive interviews. For this reason, it is important to clearly establish the purpose of the use of cognitive interviews in each study as well as the analysis strategy and methodology that will achieve it. In this case, the combined analysis strategy in which participants were first classified according to their responses, and then relying on the individual narratives, has allowed us to achieve a broad overview of what happened during the administration of the scale. Both approaches have allowed access to the differences in the participants in the construct and its indicators related to household type. These differences have provided explanations for the differences noted in the responses to the items of the scale. Nevertheless, among the disadvantages of this combined approach it is necessary to point out its complexity, although this complexity is what has allowed a systematic analysis to be carried out to optimize all the information provided by respondents during interviews.

The status of the cognitive interviewing as a qualitative method depends in part on the strategy for analysing the interview data. On general terms, one possible analysis strategy is to follow the model proposed by Tourangeau (1984), in which he examines in detail the respondents' question-and-answer

process according to the known four-phase "question-and-response" model: understanding and interpretation, retrieval, opinion and communication. By means of the so-called "standardized coding schemes" (Collins, 2007) it is possible to detect different types of problems and provide quantitative information about their occurrence (Rothgeb, Willis, & Forsyth, 2001). Nevertheless, the information provided by cognitive interviews can also be analysed from a more qualitative perspective. Miller (2007) proposes an 'interpretive' approach in three phases ranging from the analysis of individual questions to the comparison between the interpretations of different groups of respondents. Future research may address the benefits of each approach to respond to different needs when developing or evaluating scale items and survey questions.

It is important to note that the use of the cognitive interview in this type of study requires a rigorous pre-design in which the objective is clearly specified, which can be difficult in more exploratory research. Future research will focus on the usefulness of cognitive interviews for interpreting the results with other quantitative analyses such as "differential item functioning" studies, as well as finding the most appropriate way of combining qualitative and quantitative evidence to increase the quality of the measurements.

Reflections on the use of the CIRF

The CIRF is intended to help cognitive interviewers report cognitive interview studies. The CIRF has been developed thinking of the most common "scenario" in which cognitive interviewing is performed: testing survey questions to improve survey data quality. As we performed cognitive interviewing to get validity evidence for a psychological scale and provide useful insights for interpreting psychometrics, our research could be a good opportunity to test the applicability of the CIRF to other research contexts.

On general terms, the CIRF was very helpful in reporting our studies. When we performed cognitive interviewing in the Study 2, the CIRF obliged us to report on methodological issues that are often missed in cognitive interviewing

studies. For example, in the method section for Study 2 we reported demographics such as “marital status” and “type of household” and the reason why they were relevant for the study aims. In addition, we have provided a much more detailed description of the messages given to the participants for motivating them thanks to the contents of the “Ethic and data collection” item. In summary, the CIRF has made easier assess the report of Study 2. Even though Study 1 is a quantitative study, the CIRF has also helped to improve the report. The “research objective” for Study 1 presents a much more detailed argument for the need of the study than usual in the psychometric field.

Maybe the main “con” of the CIRF comes from its main focus on the use of cognitive interviewing to test survey questions. Almost all items and recommendations attend to important issue in that context, for instance: the emphasis on the phase in which cognitive interviewing is conducted with respect to the survey process, the need of identifying the target population when reporting the participant selection, or the comment on the cycled process for performing cognitive interviewing. Nevertheless, the cognitive interviewing is so flexible that can provide qualitative data for different purposes and research contexts. There is no doubt that cognitive interviewing can be a very promising within a “mixed-method” approach to social science research by itself or in combination with other methods. The CIRF authors can consider extending the framework into other methodological contexts in future updates.

Finally, we note that our explicit organization did not follow exactly that prescribed by the CIRF, as we decided to present information in a somewhat different order. We believe that report writers of cognitive interview results should be allowed the latitude to present information according to an organization that best suits the particular study.

References

- American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.
- Austin, J.K., & Huberty, T.J. (1989). Revision of the family APGAR for use by 8-year-olds. *Family Systems medicine*, 7 (3), 323-327.
- Beatty, P. & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71 (2), 287-311.
- Bellón, J. A., Delgado, J., Luna, P., & Lardelli, P. (1996). Validez y fiabilidad del cuestionario de función familiar Apgar-familiar. *Atención Primaria*, 18 (6), 289-296.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. London: Sage.
- Castillo, M., Padilla, J. L., Gómez-Benito, J., & Andrés, A. (2010). A productivity map of cognitive pre-test methods for improving survey questions. *Psicothema*, 22, 482-488.
- Collins, D. (2007). *Analysing and interpreting cognitive interview data: a qualitative approach*. Proceedings of QUEST 2007 meeting.
- Conrad, F.G., Blair, J., & Tracy, E. (1999). *Verbal Reports are Data! A Theoretical Approach to Cognitive Interviews"*. Proceedings of the Federal Committee on Statistical Methodology Research Conference
- Conrad, F.G., & Blair, J. (2009). Sources of error in cognitive interviews. *Public Opinion Quarterly*, 73 (1), 32-55.
- DeWalt, D.A., Rothrock, N., Yount, S., & Stone, A.A. (2007). Evaluation of Item Candidates The PROMIS Qualitative Item Review. *Medical Care*, 45, S12-S21.

- Feldt, L. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, 34, 363–373.
- Hardison, H.G., & Neimeyer, R.A. (2007). Numbers and narratives: quantitative and qualitative convergence across constructivist assessments. *Journal of Constructivist Psychology*, 20, 285–308
- Spanish Ministry of Health and Consume (2006). *Encuesta Nacional de Salud de España 2006* [National Health Survey in Spain 2006]. Madrid: Ministerio de Sanidad y Consumo.
- Knafl, K., Deatrck, J., Gallo, A., Holcombe, G. Bakitas, M., Dixon, J., & Grey, M. (2007). Focus on Research Methods: The Analysis and Interpretation of Cognitive Interviews for Instrument Development. *Research in Nursing & Health*, 30, 224–234.
- Mauricio, J., Romero, N., Saa, H.A., Herrera, J.A. & Reyes-Ortiz, C.A. (2006). Prevalence of religiosity, family function, social support and depressive symptoms in old people. *Revista Colombia Médica*, 37 (2 Sup. 1), 26-30.
- Miller, K. (2007) *Design and Analysis of Cognitive Interviews for Cross-National Testing*. 2007 European Survey Research Association Annual Meeting. Prague, Czechoslovakia
- Montecinos, J. (2007). Instrumentos del medico de familia en la consulta de atención primaria. *Revista Médica La Paz*, 5 (2), 63-67.
- Olt, H., Jirwe, M., Gustavsson, P., & Azita, E. (2010). Cultural Competence Among Healthcare Professionals_Revised (IAPCC-R). *Journal of Transcultural Nursing*, 21, 55.
- Poole, H.M., Murphy, P., & Nurmikko, T.J. (2009). Development and Preliminary Validation of the NePIQoL: A Quality-of-Life Measure for Neuropathic Pain. *Journal of Pain and Symptom Management*, 37 (2).

- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics*, 4 (3), 207-230.
- Rothgeb, J., Willis, G., & Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organizations Produce Similar Results?. *Proceeding of the Annual Meeting of the American Statistical Association*.
- Sarkisian, C.A., Hays, R.D., Berry, S., & Mangione, C.M. (2002). Development, reliability, and validity of the expectations regarding aging ERA-38 survey. *The Gerontologist*, 42 (4), 534.
- Smilkstein, G. (1978). The Family APGAR: A proposal for family function test and its use by physicians. *Journal of Family Practice*, 6(6), 1231-1239.
- SPSS, Inc. 2007. SPSS-16 User's guide. Chicago, USA.
- Tourangeau, R. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research*, 15, 3-7.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.
- Wolnitzki, L., Vargas, N., Cerón, C., Errázuriz, G., Fabres, J., Morales, J.C., Sepúlveda, P., Sepúlveda, X., Silva, A. & Varela, P. (1989). Social net in support of adolescents scholars: its appraisal by Smilkstein Apgar test. *Boletín Hospital San Juan de Dios*, 36 (3), 147-155.

CHAPTER 4:

A two method-two effect size measure strategy for analysing polytomous Differential Item Functioning: An illustration with Differential Step Functioning and Ordinal Logistic Regression

Benítez, I., Padilla, J.L., Hidalg, M.D, & Sireci, S.G. (Submitted). A two method-two effect size measure strategy for analysing polytomous Differential Item Functioning: An illustration with Differential Step Functioning and Ordinal Logistic Regression. *Language Testing*.

Abstract

Over the last few decades, Differential Item Functioning (DIF) has received increased attention by professionals and researchers interested in international and cross-lingual assessments. DIF analyses can provide validity evidence of the equivalence level reached by different linguistic versions of scale items. Few studies have researched ways of overcoming difficulties in detecting and explaining polytomous DIF in empirical studies for cross-lingual comparisons, such as obtaining wrong identifications and false negatives, and identifying DIF causes. The aim of the study was to illustrate how to increase confidence in DIF empirical results and obtain valuable insights into the causes of DIF in polytomous DIF by combining two statistical methods and two effect size measures. Differential Step Functioning (DSF) and Ordinal Logistic Regression (OLR) were applied along with two effect size measures to detect DIF across English and Spanish versions of seven scales included in the Student Questionnaire of the Program for International Student Assessment (OECD, 2006). The benefits from DSF results helping to guide the search for DIF causes were also discussed.

Keywords: Cross-lingual assessments, Differential Item Functioning, Differential Step Functioning, Ordinal Logistic Regression, Effect size measures, Program for International Student Assessment.

Introduction

Cross-lingual testing has become one of the most important topics in educational and psychological research. Projects like the Programme for the International Assessment of Adult Competencies, PIAAC (Organization for Economic Co-operation and Development - OECD, 2004) and the Program for International Student Assessment, PISA (OECD, 2010), are just two examples of international programs that regularly evaluate and compare people around the world. The results of such comparisons are typically disseminated in the form of rank-ordering countries with respect to educational achievement or other educational or psychological variables. Despite the fact that important decisions can be made based on comparing countries, such inferences are sometimes drawn without taking potential biases into account, like linguistic and cultural differences that can threaten the validity of rank-orderings and cross-country comparisons.

As the International Test Commission (ITC) has pointed out, when translated versions of assessment instruments are used to compare groups and individuals, the consistency of measurement across languages must be established in advance (International Test Commission, 2010). The ITC makes this clear in their *Guidelines for translating and adapting tests* by stating, specifically in D7 and D9,

Test developers/publishers should apply appropriate statistical techniques to (1) establish the equivalence of the different versions of the test or instrument, and (2) identify problematic components or aspects of the test or instrument which may be inadequate to one or more of the intended populations...Test developers/publishers should [also] provide statistical evidence of the equivalence of questions for all intended populations. (pp. 2-3)

Among the types of statistical evidences, Differential Item Functioning (DIF) analysis is useful to evaluate the equivalence level reached by different

linguistic versions of scale items. DIF occurs when examinees with the same proficiency level on the characteristic or attribute measured, but who belong to different groups (i.e., demographic, linguistic, national or cultural), have a different probability of giving a specific item response (Millsap & Everson, 1993). If individuals from different groups who are thought to be the same on the attribute measured (based on total test score or some other matching variable), respond differently to an item, the item is said to “function differentially” across the groups. DIF analyses identify items that function differentially so that these items can later be inspected to determine whether the difference may be due to some form of construct-irrelevant variance. In the case of cross-lingual assessment, one potential source of DIF is translation problems. The item meaning along with its psychometric characteristics could be altered through the translation process (Sireci, Patsula, & Hambleton, 2005; Allalouf, 2003). Thus, DIF analysis is an important tool for evaluating validity in the adaptation process for cross-lingual and cultural assessments.

DIF research has traditionally been focused on subgroups of examinees who take a test in a single language. Examples of these “monolingual” comparisons include men and women, or different cultural and ethnic groups such as African-American and Euro-American examinees. In the monolingual comparisons, the different groups respond to the same linguistic version of the test or questionnaire. More recently, DIF has been extended to cross-lingual comparisons in which the groups respond to *different* linguistic versions of tests, making the assumptions underlying the analyses less tenable (Sireci, 2005). To find a common valid matching criterion can be much more difficult when different people respond to different linguistic versions of the scale items, increasing uncertainty about DIF empirical results provided by DIF statistics (Sireci, 1997).

There are a wide variety of statistical techniques for evaluating DIF especially in dichotomous items (Hidalgo & Gómez-Benito, 2010; Millsap & Everson, 1993; Penfield, 2010; Potenza & Dorans, 1995). However, uncertainty

about DIF empirical results and how to reduce it has been much less addressed than conditions under which these statistical techniques work in simulation studies. There is a broad consensus about the general recommendation to practitioners on using more than one statistical method when detecting DIF in operational settings when trying to avoid flagging DIF false positive (Hambleton, 2006; Hidalgo & Gomez-Benito, 2010). Nevertheless, given that every DIF statistic can detect false positives, using more than one statistical method can still not be enough. On the other hand, effect size measures have proved their usefulness for making informed decisions on DIF empirical results (Zumbo, 1999, DeMars, 2011). The combination of a statistical test and an effect size measure help in reducing false identification rates (Jodoin & Gierl, 2001; French & Maller, 2007; Hidalgo & Gomez-Benito, 2010). Our methodological proposal is to apply two different methods in combination with two effect size measures then to compare the results in order to increase confidence in DIF empirical results when analyzing polytomous DIF.

On the other hand, even though DIF research has grown at the same time as interest in adapting tests and questionnaires, evidence on the causes of cross-lingual DIF are still evasive (Allalouf, Hambleton, & Sireci, 1999; Ferne & Rupp, 2007; Sireci, 1997). The inability of traditional methods for providing information concerning which score levels are involved in DIF effects can partially explain the scant results on DIF causes (Penfield, Gattamorta, & Childs, 2009). There is a clear need for statistics that not only identify items showing DIF, but also help in discovering the nature and location of DIF effect (Penfield, Gattamorta, & Childs, 2009). Such statistics would be a worthy tool for context experts when investigating whether DIF effects can be attributed to characteristics of item stems, task, or, in the polytomous case, score levels.

To implement the methodological proposal of this study, Differential Steps Functioning (DSF) statistics via DIFAS procedure (Penfield, 2005), and Ordinal Logistic Regression (OLR, Miller & Spray, 1993) were applied along with two effect size measures to detect polytomous DIF across English and Spanish

versions of seven scales included in the 2006 Program for International Student Assessment Student Questionnaire of the (OECD, 2006). Both methods and effect size measures are introduced in the following section.

Statistical procedures: Differential Step Functioning and Ordinal Logistic Regression

The Penfield's *differential step functioning* (DSF) framework is intended to detect measure invariance within each step underlying the polytomous response variable (Penfield, 2005, 2010; Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009). In contrast with traditional measures of DIF for polytomous DIF which provide only an item-level index of DIF effects, DSF allows researchers to identify the score levels involved in DIF effects. Within DSF framework, three general approaches for evaluating DSF can be applied: an IRT approach, an odds ratio approach, and a logistic regression approach. The odds ratio approach to test the null hypothesis of no DSF involves comparing the odds of successfully being in the j^{th} response category or higher on an item across examinees in different groups who are matched on the construct being measured. Specifically, the ratio of the odds of success of a reference group over the odds of success of a focal group is compared by the step-level log-odds ratio estimator ($\hat{\lambda}_j$). The natural logarithm of this common odds ratio taken across all levels of total test score is represented by the following test statistic:

$$z(\hat{\lambda}_j) = \frac{\hat{\lambda}_j}{SE(\hat{\lambda}_j)} \quad [1],$$

where $\hat{\lambda}_j$ is natural log of the common odds ratio, and j represents a single response category or score point for the item. A value of $\lambda_j = 0$ corresponds to no DSF at the j^{th} step, a value of $\lambda_j > 0$ corresponds to DSF favoring the reference group for the j^{th} step, and a value of $\lambda_j < 0$ corresponds to DSF favoring the focal group for the j^{th} step. The common log-odds ratio for the j^{th}

step can be estimated using the common log-odds ratio DSF effect estimator described by Penfield (2007) and implemented in DIFAS (Penfield, Alvarez, & Lee, 2009). The resulting estimator, $\hat{\lambda}_j$, is analogous to the Mantel-Haenszel common log-odds ratio estimator (Mantel & Haenszel, 1959) widely used in the assessment of DIF in dichotomous items (Holland & Thayer, 1988). Penfield (2007) showed that the power of DSF can be more than 10 times that of the “omnibus test”, that is, the only item-level index of DIF in the extreme case when the sign of DSF effects changes across the steps.

On the other hand, Logistic Regression provided a common framework for analyzing DIF (Zumbo, 1999). When items have more than two ordered response categories (k categories) Ordinal Logistic Regression (OLR), proposed by Miller and Spray (1993) can be used. This technique estimates a single common odds ratio assuming that the odds are proportional across all categories (Hidalgo & Gómez-Benito, 2003; Zumbo, 1999). The OLR model can be written as:

$$\ln \left[\frac{\Pr(Y \leq k | G, X)}{1 - \Pr(Y \leq k | G, X)} \right] = \beta_{0k} + \beta_1 X + \beta_2 G + \beta_3 GX$$

where $\Pr(Y \leq k)$ is the probability of response in category k or below and β_{0k} , $\beta_1, \beta_2, \beta_3$ are constants usually estimated by maximum likelihood. This model requires a separate intercept parameter for each cumulative probability. Under that formulation an item shows uniform DIF if $\beta_2 = 0$ and $\beta_3 = 0$, and a non-uniform DIF if $\beta_3 \neq 0$. These hypotheses can be tested using G^2 likelihood ratio statistic. The hypotheses about β parameters are normally tested using a conditional likelihood ratio test. For example, the likelihood ratio statistic estimated in the absence of DIF – the null model with total score only, is compared with that obtained when the model is adjusted for the presence of DIF – the full model with group, total score, and the interaction. If the difference between the two statistics is significant, the item is detected as showing DIF (this difference is assessed using a chi-square distribution with two degrees of freedom). If the effect of the group variable is significant but that

of the interaction is not, the item shows uniform DIF; if the observed test score and group interaction (GX) is significant the item shows non-uniform DIF.

Few studies have provided data about the power and Type I error rate when applying OLR. The type I error rate has been considered adequate in no DIF situations with small sample sizes (500 participants per group), and different test length (Scott et al. 2009). Power of OLR involves more than 200 participants per group except for tests or scales with only two items.

With the aim of illustrating the application of the methodological proposal, seven scales included in the PISA Student Questionnaire (OECD, 2006), were analysed by the odd-ratio approach to DSF and OLR. Following Penfield's terminology, the convergence between a "net test" of DIF (the odd-ratio approach to DSF), and a "global test" (OLR) was explored and the indications provided by DSF on level scores involved in DIF effect analysed. The scales were selected for constituting an international evaluation in which different languages adaptations are administrated. The questions inquired about students' general and personal value of science, as well as their interest and enjoyment of science, plus their self-concept of their own scientific abilities and whether they are motivated to use science in the future.

Method

Participants

Data were obtained from the PISA database (OECD, 2006), in which responses of 17,405 participants from Spain and 4,902 participants from the United States (US) were coded. DIF analyses were done using country as the group variable. With the aim of having comparable group sizes and confirmation for any statistical conclusions, two random subsamples of 2,450 participants were selected from each country giving two comparisons of 4,900 participants: Spain 1-US 1 and Spain 2-US 2. The participants from Spain were all 16 years old and those from the US were between 15 and 16 years old (Mean

15.5 and Standard Deviation 0.5). For purposes of this study, Spain was considered the reference group.

Instruments

Seven scales were selected for analysis from the 2006 PISA Student Questionnaire. All were four-point Likert item scales intended to measure science related attitudes. Table 1 presents the intended construct for each scale, the abbreviation which will be used across the paper for identifying the items in each scale and the number of items in each scale.

Table 1. PISA scales used in the DIF analyses.

Construct	Abbreviations	# of Items
Enjoyment of Science	Enj	5
General Value of Science	Gen	5
Personal Value of Science	Per	5
Instrumental Motivation to learn Science	Ins	4
Future-Oriented to Science Motivation	Fut	5
Science Self-Efficacy	Eff	8
Science Self-Concept	Con	6

These scales were selected based on two criteria. First, the scales should have been developed as unidimensional scales; and secondly, PISA researchers should use the total scale score in a descriptive manner (e.g., to describe student attitudes to science) or statistical (e.g., in the computation of students' plausible value scores). Specifically, results from these scales are used to compute psychometric analyses (e.g., computing plausible values for examinees), in statistical analyses related to test performance (e.g., in reporting correlations among test scores and interest in science, etc.), and in reporting the results.

The 2006 PISA database

Data were obtained from the OECD website (OECD, 2006). Scales and subjects were selected and data cleaning was conducted by eliminating the subjects with incomplete responses to one or more questions. This process resulted in a relatively minor loss of data (12.6% and 11.2% for the USA and Spanish samples, respectively). Subsamples were obtained from this cleaned database. Finally, DIF analyses were computed and the reliability of the results was checked by comparing the results across two independent comparisons.

Analyses

Polytomous DIF was analysed using Penfield's *Differential Step Functioning* (DSF) framework (Penfield, 2005, 2010; Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009). To conduct the DSF analyses, we used the DIFAS 4.0 software (Penfield, 2005), which evaluates DIF/DSF using the odds ratio approach to test the null hypothesis of no DSF. We used DIFAS to first analyse overall DIF (i.e., DIF at the item level), and then subsequently to evaluate DSF in items that were flagged as showing overall DIF. The Standardized Liu-Agresti Cumulative Common Log-Odds Ratio (LOR Z) was used to flag items for DIF, in which a value greater than 2.0 or lower than -2.0 is considered evidence of the presence of DIF (Penfield & Algina, 2003). DSF analysis was applied for items flagged with DIF in both subsamples applying cumulative categories with three steps (since there were four response categories in each item). The effect size for evaluating DIF item components was $\hat{\lambda}_j$ (the step-level log-odds ratio estimator), with $|\hat{\lambda}_j| < .43$ signifying a small or negligible effect, $.43 \leq |\hat{\lambda}_j| < .64$ signifying a medium effect, and $|\hat{\lambda}_j| \geq .64$ signifying a large effect (Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009). Effect size was evaluated in each step giving three effect size values for each item. The highest value in each item was extracted to be compared with OLR results in which statistical categories analyses was not possible. Lastly, the taxonomy proposed by Penfield, Alvarez and Lee (2009) based on DSF patterns, was used to interpret DIF effects.

On the other hand, OLR analyses were performed with the Statistical Package for Social Sciences (SPSS v.16) by following the instructions elaborated by Zumbo (1999). First, items were analysed in both comparisons obtaining a chi-square significance value which was used for determining items with DIF. Later, the step-level log-odds ratio estimator values and DELTA index values were checked in items detected with DIF in both samples, and DIF effect size was classified. The effect size classification was done by following the criterion explained above for the step-level log-odds ratio estimator (Penfield, Alvarez, & Lee, 2009; Penfield, Gattamorta, & Childs, 2009) and using the Educational Testing Service (ETS) criterion for interpreting DELTA index (Zieky, 1993). In this case DIF was considered as medium when values ranged between 1 and 1.5 and large when values were higher than 1.5.

Results

Descriptive Statistics

The mean (M), the standard deviation (SD), and the item-scale corrected correlation taken for item discrimination (DI) values for the items are presented in Table 2, along with the coefficient alpha reliability estimate (α) calculated for each scale. Separate statistics are presented for each sample. Both Spanish groups had relatively higher means on the majority of the items although the differences generally appear small. The SD were also similar in the four samples, with the lowest values in the General Value of Science scale. The General Value of Science scale had the lowest DI values, but item 1 in the Personal Value of Science scale, and so also the lowest levels of coefficient alpha for both country samples. Considering the habitual criterion for evaluating the items quality, all the scales except General Value of Science, obtained adequate results, that's, means were around the central point of the scales, and SD and DI reached values highest than 0.5 in all the samples.

Table 2. Descriptive statistics for the PISA scale items

Item		Spain 1			US 1			Spain 2			US 2		
		M	SD	DI	M	SD	DI	M	SD	DI	M	SD	DI
Enjoyment of Science	1	2.39	.86	.82	2.33	.81	.83	2.42	.85	.81	2.33	.83	.83
	2	2.60	.82	.78	2.54	.82	.81	2.61	.82	.78	2.55	.83	.81
	3	2.88	.81	.66	2.63	.80	.78	2.89	.82	.67	2.64	.81	.79
	4	2.32	.84	.80	2.24	.81	.83	2.36	.86	.81	2.24	.82	.83
	5	2.24	.84	.81	2.26	.86	.84	2.27	.86	.81	2.26	.86	.85
Scale α		.91			.93			.91			.93		
Science Self-Efficacy	1	2.31	.85	.53	1.96	.78	.77	2.32	.85	.52	1.96	.76	.58
	2	1.95	.87	.56	1.92	.86	.86	1.97	.89	.56	1.94	.87	.60
	3	2.39	.96	.56	2.23	.90	.90	2.44	.98	.57	2.24	.90	.66
	4	2.42	.90	.59	2.23	.86	.86	2.43	.92	.59	2.24	.84	.61
	5	2.28	.97	.62	1.90	.86	.86	2.29	.97	.61	1.94	.86	.62
	6	2.22	.93	.53	2.07	.87	.86	2.26	.93	.51	2.07	.85	.61
	7	2.43	.97	.55	2.33	.93	.92	2.43	1.01	.53	2.35	.92	.62
	8	2.25	.96	.56	2.36	.95	.93	2.24	.97	.56	2.36	.93	.63
Scale α		.83			.87			.83			.87		
General Value of Science	1	1.41	.55	.50	1.68	.65	.65	1.43	.60	.52	1.66	.64	.62
	2	1.56	.60	.45	1.61	.62	.62	1.57	.61	.51	1.61	.61	.61
	3	1.93	.73	.43	1.85	.68	.67	1.93	.74	.46	1.83	.67	.60
	4	1.83	.66	.49	1.76	.67	.66	1.85	.67	.50	1.76	.66	.64
	5	1.80	.67	.55	2.04	.76	.74	1.81	.69	.56	2.04	.73	.56
Scale α		.72			.82			.74			.82		
Personal Value of Science	1	2.30	.77	.36	2.12	.73	.52	2.36	.78	.36	2.11	.74	.52
	2	2.20	.77	.65	2.10	.80	.71	2.21	.78	.66	2.08	.80	.72
	3	2.35	.84	.68	2.20	.81	.73	2.37	.83	.69	2.19	.82	.73
	4	2.01	.72	.59	1.97	.74	.67	2.03	.74	.60	1.96	.74	.68
	5	2.30	.85	.62	2.09	.80	.67	2.30	.85	.62	2.06	.80	.67
Scale α		.80			.85			.79			.85		
Future-Oriented to Science Motivation	1	2.62	1.03	.87	2.56	.95	.84	2.65	1.04	.87	2.57	.94	.83
	2	2.62	1.04	.85	2.58	.94	.84	2.66	1.05	.84	2.58	.94	.84
	3	3.02	.90	.83	2.95	.86	.82	3.05	.86	.83	2.95	.87	.83
	4	2.95	.91	.81	2.85	.87	.74	2.97	.91	.80	2.88	.87	.73
Scale α		.93			.92			.93			.91		
Instrumental Motivation to learn Science	1	2.18	.96	.82	1.97	.79	.78	2.19	.97	.82	1.95	.78	.77
	2	2.33	.98	.86	2.16	.84	.83	2.37	.99	.85	2.14	.84	.81
	3	2.21	.90	.79	2.03	.77	.76	2.26	.91	.78	2.03	.77	.77
	4	2.25	.91	.86	2.11	.82	.82	2.27	.93	.85	2.11	.83	.82
	5	2.27	.91	.84	2.12	.83	.82	2.32	.92	.81	2.12	.82	.81
Scale α		.94			.92			.93			.92		
Science Self-Concept	1	2.43	.79	.66	2.36	.84	.84	2.45	.80	.68	2.34	.83	.83
	2	2.33	.78	.75	2.23	.78	.79	2.34	.78	.74	2.22	.78	.81
	3	2.44	.80	.81	2.26	.79	.79	2.46	.80	.81	2.25	.79	.83
	4	2.45	.83	.84	2.45	.82	.83	2.48	.83	.82	2.44	.82	.74
	5	2.42	.80	.80	2.18	.74	.75	2.47	.81	.79	2.18	.74	.74
	6	2.47	.79	.80	2.30	.82	.82	2.50	.80	.80	2.31	.81	.84
Scale α		.92			.93			.92			.93		

The dimensionality of responses to items in both country samples was analysed using exploratory factor analysis (principal axis method). Separate analyses were done for each scale. In all the country samples a dominant single factor was obtained, with the first factor accounting for at least 46% of the variance in the data for all the scales. The explained variance percentage ranged from 46% for the Science Self-Efficacy scale in Spain 2, to 80% for the Instrumental Motivation to Learn Science scale in Spain 1. These values confirm the unidimensionality of the scales according to usual criteria established in the literature (Carmines & Zeller, 1979; Reckase, 1979).

DIF results: Differential Step Functioning

DIF analyses were computed for all 38 items across the seven selected scales. The analyses were done separately for each scale, thus, the matching criterion was the total score across the subset of items comprising each scale. Following the criterion for the Standardized Liu-Agresti Cumulative Common Log-Odds Ratio (LOR Z), 29 items in the comparison Spain 1-US 1 and 27 items in the comparison Spain 2-US 2 were flagged with DIF. DSF was then applied to these items and effect size measure was obtained. These results are summarized in Table 3. Table 3 indicates whether the item was flagged for overall DIF in each comparison and the highest effect size estimated in one of two comparisons. The amount of DIF varied across the scales, ranging from no substantive DIF (for the two motivation scales) to all items flagged with substantive DIF (General Value of Science scale).

Table 3. Summary of DSF Results

Scale	Item	Comparison			
		Spain 1- US 1		Spain 2 - US2	
		DIF	Effect	DIF	Effect
Enjoyment of Science	1	Y	L	N	
	2	Y	S	Y	M
	3*	Y	L	Y	L
	4	N		Y	L
	5*	Y	L	Y	L
Science Self-Efficacy	1*	Y	L	Y	L
	2*	Y	L	Y	M
	3	N		N	
	4	N		N	
	5*	Y	L	Y	L
	6	N		N	
	7	Y	S	Y	L
	8*	Y	L	Y	L
General Value of Science	1*	Y	L	Y	L
	2*	Y	L	Y	M
	3*	Y	L	Y	L
	4*	Y	L	Y	L
	5*	Y	L	Y	L
Personal Value of Science	1*	Y	L	Y	M
	2	Y	S	Y	S
	3	N		N	
	4*	Y	M	Y	M
	5*	Y	L	Y	L
Future-Oriented to Science Motivation	1	N		N	
	2	Y	L	N	
	3	Y	S	Y	S
	4	Y	M	N	
Instrumental Motivation to learn Science	1*	Y	L	Y	L
	2	N		N	
	3	N		Y	S
	4*	Y	M	Y	M
	5	Y	L	Y	S
Science Self-Concept	1*	Y	L	Y	M
	2	N		N	
	3*	Y	M	Y	M
	4*	Y	L	Y	L
	5*	Y	L	Y	L
	6	Y	M	N	

DIF classification: Y=flagged for DIF, N=not flagged. Effect lists largest effect size for flagged step with S=small, M=medium, and L=Large

As Table 3 shows, 20 items were marked as exhibiting in samples either medium or large DIF. Of them, 17 had the same effect classifications in both comparisons (considering only medium and large effects) and three of them (Eff2, Gen2 and Per1) registered different effect sizes in both cases. Only seven items were not flagged with DIF in any of the comparisons. In some cases incongruous results were found, as for example Enj1 or Enj4, which were flagged with large DIF in one of the comparisons but did not exhibit DIF in the other one. In relation to the effect size results, the table shows that most items reach the highest classification at least in one of the steps. In some scales, all the steps are classified as having large DIF and in others only one or two steps which are frequently steps two or three, showing higher DIF affects in the upper categories. Additional information on effect size magnitude and sign can be found for the items flagged with large DIF by the two statistical methods in the following sections.

DIF results: Ordinal Logistic Regression

DIF analyses were computed by replicating the application conditions explained above for DSF. First, items flagged with DIF were selected by considering the chi-squared significance. Later, two criteria were applied for evaluating the effect size in only items flagged in both samples. The cut scores proposed by Penfield, Gattamorta, and Childs (2009) to assess the step-level log-odds ratio estimator were also used for the log-odds ratio estimator so that DSF and OLR results were comparable. On the other hand, a new criterion based on ETS rules was added (Zieky, 1993). In this case, DELTA index which is obtained with the step-level log-odds ratio estimator transformation classifying DIF as medium when values between 1 and 1.5 and large when values higher than 1.5. Table 4 shows items flagged with DIF in each comparison and the classification of effect size measures considering both criteria described above.

Table 4. Summary of OLR results

Scale	Item	Comparison					
		Spain 1 -US 1			Spain 2 -US 2		
		DIF	Log-odds ratio estimator	Effect ETS criterion	DIF	Log-odds ratio estimator	Effect ETS criterion
Enjoyment of Science	1	Y	L	L	N		
	2*	Y	L	L	Y	L	L
	3*	Y	M	M	Y	M	M
	4	N			N		
	5*	Y	L	L	Y	L	L
Science Self-Efficacy	1*	Y	M	M	Y	M	M
	2*	Y	L	L	Y	L	L
	3	N			N		
	4	N			N		
	5*	Y	M	M	Y	M	M
	6	N			N		
	7*	Y	L	L	Y	L	L
	8*	Y	L	L	Y	L	L
General Value of Science	1*	Y	L	L	Y	L	L
	2	Y	L	L	N		
	3*	Y	M	M	Y	M	M
	4	Y	M	M	Y	M	S
	5*	Y	L	L	Y	L	L
Personal Value of Science	1*	Y	L	L	Y	L	L
	2*	Y	L	L	Y	L	L
	3	N			N		
	4*	Y	L	L	Y	L	L
	5*	Y	L	L	Y	L	L
Future-Oriented to Science	1	N			N		
	2	Y	L	L	N		
	3*	Y	L	L	Y	L	L
	4	Y	L	L	N		
Instrumental Motivation to learn science	1*	Y	L	L	Y	L	L
	2	N			Y	L	L
	3	N			Y	L	L
	4*	Y	L	L	Y	L	L
	5*	Y	L	L	Y	L	L
Science Self-Concept	1*	Y	L	L	Y	L	L
	2*	Y	L	L	Y	L	L
	3*	Y	L	L	Y	L	L
	4*	Y	L	L	Y	L	L
	5	Y	M	M	Y	M	S
	6*	Y	L	L	Y	L	L

DIF classification: Y=flagged for DIF, N=not flagged. Effect lists largest effect size for flagged step with S=small, M=medium, and L=Large

As Table 4 shows, 24 items were detected as exhibiting medium or large DIF considering both criteria for both comparisons. In all the items, the effect size estimated was consistent. Inconsistencies were only found for two items (Gen4 and Con5), for which a medium effect size was estimated for all comparisons and criteria except DELTA index in Spain 2 -US 2

DIF results: Convergence across methods

To analyse convergence across methods, only items flagged with medium or large DIF in both comparisons were considered in order to reduce uncertainty of DIF empirical results. Table 5 shows the items flagged with DIF by two methods for both comparisons showing DIF in the same magnitude. Items are identified by the scale abbreviations specified in Table 1 and the item number in the scale.

Table 5. Convergence across methods

		OLR	
		Large	Medium
DSF	Large	Enj5, Eff8, Gen1, Gen3, Gen5, Per5, Ins1, Con4	Enj3, Eff1, Eff5
	Medium	Per4, Ins4, Con3	

Enj=Enjoyment of Science; Gen=General Value of Science; Per=Personal Value of Science; Ins=Instrumental Motivation to learn Science; Eff= Science Self-Efficacy; Con= Science Self-Concept.

Table 5 shows the convergence reached by the two statistical methods and the effect size measures. There was a complete agreement on eight items across methods and comparisons. The eight items were those classified as having large DIF in both analyses. On the other hand, six items were flagged by both methods but with different effect sizes. Another three items which are not included in the table were flagged with medium or large DIF but differences in one comparison were found for one of the methods. For example, item Eff2 was flagged with medium DIF by OLR in both comparisons but only in one by DSF.

DSF analyses for DIF items

A deep review of the DSF was done to obtain insights into the nature of the nature and location of the DIF effect for the eight items flagged with large DIF for both methods in the two comparisons. First, we identified the step or steps involving DIF for each item in both comparisons. Secondly, the magnitude of DSF was determined following the indications of the taxonomy developed by Penfield, Alvarez and Lee (2009). Lastly, the sign of DSF and suggestions on the characteristics of the item -item stem, task, or score level-, potentially biased were added. Table 6 summarizes the review of the DSF analyses for the eight items on which both methods converge across comparisons.

Table 6. DSF steps, magnitude and sign in items with DIF.

Item	Spain 1- US 1					Spain 2- US 2				
	Step	Mag	Sig	Form	Revision	Step	Mag.	Sig	Form	Revision
Enj5	2,3	L	US	NP-Cs	SL	1 2,3	M L	US	P-Co	IL
Eff8	1,2,3	L	US	P-Cs	IL	1,2,3	L	US	P-Cs	IL
Gen1	1,2,3	L	US	P-Cs	IL	1 2,3	L M	US	P-Co	IL
Gen3	2,3	L	Sp	NP-Cs	SL	1 2,3	M L	Sp US	P- D	IL
Gen5	1,2 3	L M	US	P-Co	IL	1,2 3	L M	US	P-Co	IL
Per5	3	L	Sp	NP-Cs	SL	3	L	Sp	NP-Cs	SL
Ins1	2	M	Sp	NP-Cs	SL	2	M	Sp	NP-Cs	SL
Con4	1,2 3	L M	US	P-Co	IL	1,2	L	US	NP-Cs	SL

DSF magnitude: M=Medium, and L=Large; DSF form: P=Pervasive, NP= Non-Pervasive, Cs= Constant, Co= Convergent and D= Divergent; Level of revision: IL= item level and SL= Score level

Three items showed pervasive DSF (Eff8, Gen1, Gen5) as all steps displayed a substantial DSF against the US samples across comparisons. The signs of DSF

effect for these three items mean that the advance from the lowest score levels to a higher score level is easier for the Spanish respondents than for the US respondents. The DSF of one of these items (Eff8) was constant, the sign and magnitude being equal across steps. Following advice by Penfield, Alvarez and Lee (2009), context experts should look at the item stem to identify DIF causes because the DIF effect is located at the item level. Item Gen5 showed a convergent DSF since the sign of DSF was equal but magnitude changed across steps; while for item Gen1 the classification of DSF varied across comparisons being constant in one of them and convergent in the other one. In addition, DSF for items Per5 and Ins1 were non-pervasive across comparisons pointing out that the DIF effect could be located at the score levels. The sign of DSF for both items means that the transition from the lowest score levels to a higher score level was relatively easier for the US participants than for the Spanish participants. Content expert should review score levels to identify DIF causes. Finally, DSF for three items was classified differently across comparisons being pervasive in one of them and non-pervasive in the other one. The sign of DSF does not change across comparisons: Spanish groups could find it easier to advance from the lowest score levels to a higher one in item (Enj5), while the opposite was found in items (Gen3, Con4). Item stems and response categories should be checked by content experts by looking for DIF causes in these three items.

Discussion

The aim of this study was to illustrate how to overcome habitual difficulties in detecting and explaining polytomous DIF in empirical studies by using a two methods-two effect size measure strategy. To illustrate the methodological proposal, data from attitudinal scales in 2006 PISA Student Questionnaire were analysed by DSF and OLR methods. Comparable effect size measures were also computed to increase confidence in the DIF empirical results. Despite the odd ratio approach being conceptually different (net test of DIF) within the DSF framework and ORL (global test of DIF), a substantive convergence of results

was found across comparisons in a cross-validation design. 14 items were flagged with DIF by both methods and criteria for the estimated effect size measures, reaching a “complete convergence” in eight items flagged with DIF in all conditions showing the same magnitude of DIF.

The “two methods-two effect size measure strategy” increased confidence in DIF empirical results. The number of items flagged with DIF decrease when looking at the convergence across methods and effect size measures. The reduction of item flagged with DIF allows researchers and content experts to focus on “evident” DIF effect instead of being distracted by “false positive” DIF. To focus on “evident” DIF effect is significantly important when DIF studies are performed during post-operational phases for international assessment projects like PISA since removing or changing item characteristics is not possible, but identifying DIF causes can be very useful to improve validity of cross-lingual and country comparisons.

With regard to the DIF methods used in this study, even though OLR only provide item-level DIF index, it allows us to estimate effect size measures comparable to those estimated for the DSF analyses when odd ratio approach is followed. In addition, a high degree of convergence was found between the net test computed within the DSF framework and the OLR results, which following Penfield (2010) terminology can be considered a global test. On the other hand, DSF framework has proved to be very powerful for analysing the nature of DIF effect, location, and guiding content experts when investigating DIF causes. The taxonomy proposed by Penfield, Alvarez, and Lee (2009) is a worthy complement to the tools available to identify DIF causes. For example, three in five items of the General Value of Science scale showed large DIF and pervasive DSF across all conditions but one (Gen3 for Spain 1-US1). This DSF patterns suggest context experts should look at the item stem to find DIF causes. The US version of the item stems for three DIF items of the General Value of Science scale include the expression “*advanced in broad science and technology*”, while the Spanish version of these three items exclude the word “broad”. Thus, difference

in the linguistic versions for both country samples could explain the DSF effect found.

Finally, mixed-method studies combining quantitative and qualitative evidence could help in increasing the knowledge of DIF causes (Padilla, et al. 2010). There is no doubt about the benefits for cross-lingual and country assessments coming from solid metric equivalence evidence and a deeper understanding of DIF causes.

References

- Allalouf, A. (2003): Revising Translated Differential Item Functioning Items as a Tool for Improving Cross-Lingual Assessment. *Applied Measurement in Education*, 16 (1), 55-73.
- Allalouf, A., Hambleton, R.K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36 (3), 185-198.
- Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. London: Sage.
- DeMars, C.E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, 24 (3), 189-209.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113-148.
- French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement* 67, 373-393.
- Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44 (11), S182-S188
- Hidalgo, M. D., & Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment* 19, 1-11.
- Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (3rd edition)*. USA: Elsevier - Science & Technology.

- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.
- International Test Commission (2010). *International Test Commission Guidelines for Translating and Adapting Tests*. Retrieved January 13, 2012, from <http://www.intestcom.org>.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with Logistic Regression procedure for DIF detection. *Applied Measurement in Education* 14, 329-349.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Miller, T.R., & Spray, J.A. (1993). Logistic Discriminant Function Analysis for DIF Identification of polytomously scored items. *Journal of Educational Measurement* 30 (2), 107-122.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 17, 297-334.
- Organisation for Economic Co-operation and Development. (2004). *Programme for the International Assessment of Adult Competencies (PIAAC)*. Policy Objectives, Strategic Options and Cost Implications. Stockholm: Author.
- Organisation for Economic Co-operation and Development. (2006). *PISA 2006 database*. Retrieved July 20, 2010, from <http://pisa2006.acer.edu.au/downloads.php>.
- Organisation for Economic Co-operation and Development. (2010). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*. Retrieved January 13, 2012, from <http://www.oecd.org/dataoecd/10/61/48852548.pdf>.

- Padilla, J.L., Benítez, I., Hidalgo, M.D., & Sireci, S.G. (2010). *Detecting sources of DIF in polytomous items by cognitive interviewing*. Paper presented at 41st Annual Conference of Northeastern Educational Research Association (NERA). Hartford, Connecticut.
- Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 40*, 353-370.
- Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.
- Penfield, R.D. (2007): An Approach for Categorizing DIF in Polytomous Items. *Applied Measurement in Education, 20* (3), 335-355.
- Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*, 129-149.
- Penfield, R.D., Alvarez, K. & Lee, O. (2009): Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF in Polytomous Items: An Illustration. *Applied Measurement in Education, 22* (1), 61-78
- Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items. *Educational Measurement: Issues and Practice, 28*, 38-49.
- Potenza, M. T., & Dorans, N. J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19*, 23-37.
- Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of educational statistics, 4* (3), 207-230.
- Scott, N.W., Fayers, P.M., Aarson, N.K., Bottomley, A., de Graeff, A., Groenveld, M., Gundy, C., Koller, M., Petersen, M.A., & Sprangers, M.A.G. (2009). A simulation study provided sample size guidance for differential

- item functioning (DIF) using short scales. *Journal of Clinical Epidemiology*, 62, 288-295.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16 (1), 12-19.
- Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). New Jersey: Lawrence Erlbaum Associates.
- Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. En R.K. Hambleton, P.F. Merenda y S.D. Spielberger (eds.): *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). New Jersey: Lawrence Erlbaum Associates.
- SPSS, Inc. 2007. SPSS-16 User's guide. Chicago, USA.
- Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.). *Differential Item Functioning* (pp. 337-347). Hillsdale, NJ: Erlbaum.

CHAPTER 5:

Analysis of the causes of DIF using a Mixed Methods approach: Analysis of Cognitive Interviews and DIF.

Benítez, I., & Padilla, J.L. (Submitted). Analysis of the causes of DIF using a Mixed Methods approach: Analysis of Cognitive Interviews and DIF. *Journal of Mixed Methods Research*.

Abstract

Differential Item Functioning (DIF) can undermine the validity of cross-lingual comparisons. While a lot of efficient statistics for detecting DIF are available, few general findings have been found to explain DIF results. The objective of the paper was to study DIF sources by using a Mixed Method design. The design involves a quantitative phase in which DIF was analysed followed by a qualitative phase conducting cognitive interviews. To illustrate the proposal, polytomous DIF was analysed in the scales from the PISA Student Questionnaire (OECD, 2006). Evidence obtained allowed DIF to be connected with differences in the interpretation patterns of participants from the different linguistic groups. Lastly, benefits of Mixed Methods design for analysing equivalence in cross-lingual assessments will be discussed.

Keywords: Cross-lingual assessments, Differential Item Functioning causes, Mixed Methods, Cognitive interviews, Program for International Student Assessment.

Introduction

Cross-lingual and cross-cultural assessment has become one of the most important topics in educational and psychological research. Some of the most important studies are the Progress in International Reading Literacy Study (PIRLS; IEA, 2011), led by the International Association for the Evaluation of Educational Achievement (IEA) to internationally assess the reading skills of children aged 9 and 10 years; the Program for International Student Assessment (PISA; OECD, 2009) coordinated by the Organization for Economic Co-operation and Development (OECD), which evaluates educational competencies and attitudes of 15 year-olds from different countries; or the Programme for the International Assessment of Adult Competencies (PIAAC; OECD, 2011) also coordinated by OECD to study adult skills and ability for solving problems in a technology-rich environment.

The comparison of the groups evaluated in the international studies has involved different language versions of tests and questionnaires. The usual practice is to create several original versions which are then adapted to other languages so participants respond in their mother tongue. Despite the rigour of the adaptation processes, the International Test Commission (2010) points out the need for an empirical assessment of the degree of equivalence between different language versions in order to ensure the validity of the inferences raised about the differences or similarities between groups, and to detect biases unrelated to the construct being measured.

So far, the methods used to study the threats to the comparability of the groups and therefore the validity of inferences, have not provided sufficient overall results. Usually statistical techniques are applied in order to learn the level of equivalence achieved between different language versions of tests and questionnaires (Sireci, Patsula, & Hambleton, 2005). Analysis of Differential Item Functioning (DIF) is one of the methods most used to determine if the level of metric equivalence has been achieved (Van de Vijver, & Poortinga, 2005). DIF occurs when people with the same level in the measured

characteristic, ie, comparable, have different probabilities of giving a particular response to an item depending on the group they belong to (Millsap, & Everson, 1993).

The DIF study has progressed in the development of statistical techniques to detect items with DIF much more than in the extending the knowledge of the factors responsible for its appearance. There is a wide variety of statistical techniques for evaluating DIF especially in dichotomous items (Hidalgo, & Gómez-Benito, 2010; Zumbo, 2009). However, even though DIF research has grown along with interest in adapting tests and questionnaires, evidence on the causes of cross-lingual DIF are still evasive (Allalouf, Hambleton, & Sireci, 1999; Ferne, & Rupp, 2007; Sireci, 1997). In addition, the inability of traditional methods to provide information concerning the nature and location of DIF effects can also partially explain the scarcity of results on DIF causes (Penfield, Gattamorta, & Childs, 2009b).

The inadequacy of the methodologies used and the complexity of the phenomenon are some of the reasons that have slowed the progress of the search for the causes of DIF. Since its inception, the study of the causes of DIF was addressed from two approaches, as noted by Schmeiser (1982): statistical methods and judgment methods. The first approach has focused on identifying items with DIF for some of the groups evaluated and comparing the different methods to determine what works best. The judgment methods, through expert review, have been used primarily to eliminate the apparent bias in language and content in the sample.

Currently, statistical methods and judgment methods continue to be applied separately, although in some cases now also combining both approaches. In the statistical approach are studies such as Allalouf, et al. (1999), who applied the Mantel Haenszel (MH) to various parallel forms of a test to locate the type of items that recorded higher amounts of DIF, or the study of Swanson, Clauser, Case, Nungester, and Featherman (2002) which used Hierarchical Logistic Regression to combine the results of the Logistic Regression applied to

individual items and identify the causes of consistent DIF across the items. Allalouf, et al. (1999) identified the differences in the difficulty of words and cultural relevance for each group as causes of DIF in item content; while Swanson, et al. (2002) found a relationship between the magnitude of DIF and item characteristics such as the inclusion of drawings.

One of the studies that combine the approaches of statistical methods and judgment methods was done by Ercikan, et al. (2010). The design was the implementation of Linn-Harnisch (L-H) method (Linn, & Harnisch, 1981) and the Simultaneous Item Bias Test (SIBTEST; Shealy, & Stout, 1993) to detect DIF, and the use of experts and Think-Aloud Protocols (TAP) to locate areas in the content or context that may cause DIF. The authors pointed out differences between groups in terms of clarity and specificity of the keywords in the wording of the items as causes of DIF. Also Elosua and Lopez-Jauregui (2007) applied both approaches to evaluate the DIF using the MH statistic and using experts to categorize the items according to their structural characteristics. After comparing the results of both methods, they classified the causes of DIF in the following categories: cultural specificity, poor translation, grammatical and semantic differences.

Roussos and Stout (1996a) took a step forward in the combination of methods proposing what they called "Multidimensionality based DIF analysis paradigm." Their proposal is based on a multidimensional model of DIF in which assumptions are made about the presence of DIF in terms of substantive characteristics of the items to assess such hypotheses using statistical techniques. Gierl and Khaliq (2001) followed this approach using experts to categorize items with DIF and applying SIBTEST to detect statistical differences between the different categories previously hypothesized.

However, despite the efforts of researchers, lack of conclusive evidence has meant that consensus on the causes of DIF has not yet been reached. The literature on the causes of DIF has demonstrated the complexity of the phenomenon and the shortcomings of both statistical methods and the

judgment methods for addressing the study separately. Also, the combination of methods has produced an insufficient increase in the results. Hence our proposal to address the study of the causes of DIF from an intended mixed perspective that combines statistical and qualitative methods to achieve, as indicated by Edmeades, et al. (2010), more adequate data for establishing the conclusions sought. This proposal is made in the framework of Mixed Research (MR), which is presented as the appropriate methodological framework when the complexity of the phenomenon requires data from different perspectives as the causes of DIF does (Clarke, & Yaros, 1988).

So far, implementation of judgment methods has focused mainly on expert appraisal. However, as is the case when evaluating questions in the surveys, the contribution of experts is limited, since they have no access to the respondents' response processes of those items being tested. The logic of this proposal is to include the people who respond to the items in order to detect group differences in the response processes of test items and questionnaires. The mixed design will allow us to learn and connect these differences with the DIF.

To investigate the response process of the participants CI have been used, as it is the method most used to obtain evidence on the respondents' question-and-answer process (Willis, 2005). The CI consists of the administration of the survey questions while collecting additional verbal information about the questions, information that is then used to evaluate the quality of responses or to determine if the information sought is being collected (Beatty, & Willis, 2007). Besides the advantages of CI have shown during the pretest of the questionnaires of the surveys (Presser, et al. 2004), in the context of cross-cultural testing CI is also presented as a method for evaluating functional equivalence, i.e. for ensuring that the instruments applied to different groups are equivalent.

The overall objective of the study is to propose an approach to the study of the causes of DIF using the framework of MR. The design of the study will be conducted by combining quantitative methods for analyzing DIF and

qualitative methods for obtaining information about the respondents' question-and-answer processes when responding to items. The integration of the results will connect the presence of DIF with the discourses of the participants, so that it is possible to explain and interpret the results and reach conclusions about the causes of DIF. To illustrate this proposal, some scales contained in the Student Questionnaire of the PISA 2006 were used, whose items were polytomous with four Likert type response alternatives and which evaluated attitudes toward science.

To carry out the study, a Mixed Method Design was developed that, as described by Tashakkori and Teddlie (1998), includes two phases in which different methods were applied on different data. In each phase of the study a research question was posed, which in the quantitative phase was "Which items contain DIF?" And in the qualitative "How do the participants of different groups answer the items with DIF?" The aim of this study is that the combination of both phases answers the general question "what could the factors responsible for the DIF be?"

Quantitative phase

Participants

To carry out the DIF analyses, data were obtained from the PISA database (OECD, 2006), in which responses of 17,405 participants from Spain and 4,902 participants from the United States (US) were coded. With the aim of having comparable group sizes and confirmation for any statistical conclusions, two random sub-samples of 2,450 participants were selected from each country giving two comparisons of 4,900 participants: Spain 1- US 1 and Spain 2- US 2. The participants from Spain were all 16 years old and those from the US were between 15 and 16 years old (Mean 15.5 and Standard Deviation 0.5). For purposes of this study, Spain was considered the reference group.

Analyses

Polytomous DIF was analysed using *Penfield's Differential Step Functioning* (DSF) framework (Penfield, 2005, 2010; Penfield, Alvarez, & Lee, 2009a; Penfield, et al., 2009b) and Ordinal Logistic Regression (OLR; Miller, & Spray, 1993). Two different methods in combination with two effect size measures were applied to compare the results in order to increase confidence in DIF empirical results when analyzing polytomous DIF.

The DSF analyses were performed using DIFAS 4.0 software (Penfield, 2005). The framework of DSF is intended to overcome the limitations of omnibus DIF statistics for measuring equivalence in relation to each score level of the polytomous item. DSF provides a mechanism for examining the between-group differences in measurement properties at each step. A step is described as the chance that an individual will progress from one score level to a higher score level which enables precise identification of which score levels (or steps) are responsible for an observed DIF effect. DSF assumes a graded response model which uses a cumulative form, because in this model the step function describes the probability that an examinee successfully advances to a score equal to or greater (Penfield, et al., 2009b). DIFAS was used to first analyze overall DIF (i.e. DIF at the item level), and then subsequently to evaluate DSF in items that were flagged as showing overall DIF. DSF analysis was applied for items flagged with DIF in both comparisons applying cumulative categories with three steps since attitudinal items in the Student Questionnaire of PISA 2006 are four-point Likert item scales. On the other hand, OLR analyses were performed with the Statistical Package for Social Sciences (SPSS v.16) by following the instructions elaborated by Zumbo (1999). OLR was analyzed in both comparisons obtaining a chi-square significance value which was used for determining items with DIF.

Next, the effect size of the DIF was assessed. The effect size classification was done by following the criterion for the step-level log-odds ratio estimator (Penfield, et al., 2009a; Penfield, et al., 2009b) and using the Educational Testing

Service (ETS) criterion for interpreting DELTA index (Zieky, 1993). The next step was to select those items classified as Large DIF by both methods and in the two comparisons. Later, a deep review of the DSF was done to obtain insights into the nature and location of the DIF effect for the eight items flagged with large DIF by both methods in the two comparisons. First, we identified the step or steps involving DIF for each item in both comparisons. Secondly, the magnitude of DSF was determined following the indications of the taxonomy developed by Penfield, et al. (2009a). The authors classified the DSF using a taxonomy with two dimensions: pervasiveness and consistency. Both dimensions categorize items based on the number of steps in which DIF appears, the magnitude and the sign of DIF, that is, taking DIF effect size and the group disadvantaged by the item into account for each of the steps with DIF. The pervasiveness dimension distinguishes between Pervasive DSF, which occurs when all steps display a substantial DSF effect; and Non-pervasive DSF which corresponds to the situation whereby some steps, but not all, display a substantial DSF effect. Furthermore, the consistency can be classified as: Constant, when DSF effects are equal in magnitude and sign across steps; Convergent, when DSF effects have the same sign, but not the same magnitude across steps; and Divergent, which occurs when DSF effects have different signs.

Qualitative phase

Participants

For the application of Cognitive Interviewing (CI) 44 participants were recruited, 24 from Spain (15 women and 9 men) and 20 of the US (11 women and 9 men). These participants were chosen to mimic the characteristics of participants in the PISA study: Students between 15 and 16 years who were in the final stages of compulsory education. Participants responded to the Student Questionnaire scales of the PISA 2006 study in their mother tongue.

The US interviews took place in Chicago and six suburbs ranging from the far south suburban Chicago area to the northern suburbs. Five US respondents

were recruited using a local youth job center, the rest through word of mouth. There was one private school student; the rest went to local public schools. Schools included a wide range of socioeconomic and ethnic diversity. Two interviews were conducted in a public library study room, four at local coffee shops and the rest in private homes. Spanish interviews were conducted in Granada. All respondents were contacted through school principals and word of mouth. There were 12 private school students, and 12 students that attended a local public school. In Spain the public school includes a wide range of socioeconomic diversity; while the private school includes middle and upper-middle socioeconomic status. During CI, participants were informed about the purpose of the study. The interviews were recorded on audio with their consent and the consent of participants' parents. The participants were guaranteed confidentiality and that the data would be solely used for purposes related to research. In addition, each participant was rewarded with a memory stick. The interviews were conducted individually by experienced interviewers who were trained to reproduce the characteristics of the original administration, ie, first the participants responded to the scales and then they were interviewed.

Materials

The interviewers used an interview protocol that included the selected items in the quantitative phase and the suggestions of a group of experts. 12 experts were asked to rate to what extent US English and Spanish items were comparable and to provide comments on linguistic issues: terms, expressions, etc. that could undermine item comparability. Expert comments about items were transformed in to follow-up probes, which were applied to discover how subjects go through the "question-and-answer" process while responding to the items. The interview protocol was translated and administered in the United States and Spain.

Analyses

Analyses were done by using transcriptions and by means of the Q-notes software (<https://wwwn.cdc.gov/qnotes/login.aspx>). Q-Notes was developed

by the National Center for Health Statistics (NCHS) and supports the structured collection and analysis of CI data. It is available for use by US government agencies and survey research organizations.

The analysis of the CI data was conducted in several stages by following the approach developed by Miller (2007). These types of analyses follow a sociological approach, which is focused on respondents' interpretations; specifically, on analyzing how respondents put their own life experience into responses to items and questions. Its aim is to understand the meanings people bring to items by investigating the respondents' interpretations of their cognitive processes.

The Miller (2007) approach provides three phases in the data analysis. The first phase involves the analysis of individual interviews. In this case the interviews were reviewed to obtain the *themes* or general topics developed by the respondents for each item. In the second stage of analysis, *subthemes* or specifications of the theme are extracted, referring for example to the type of situations narrated or the people named. Finally, interviews are coded using the themes and subthemes in order to reach the third level of analysis, comparison between groups.

In order to increase confidence in the results, four professionals participated in the process of coding the interviews, three analysts and an adjudicator. The first analyst, who was bilingual and an expert in the analysis of CI, performed the initial coding of the 44 interviews using a detailed description of each of the themes and subthemes. Then, the second analyst, a professional in the field of test and questionnaire development and cross-cultural assessment, reviewed the interviews with the Spanish participants, while the third analyst, a native English speaker, familiar with American and Spanish cultures, reviewed interviews with American participants. Both noted their disagreement with the initial coding and gave reasons for such disagreements. The first analyst resolved the discrepancies by making changes when the arguments of the

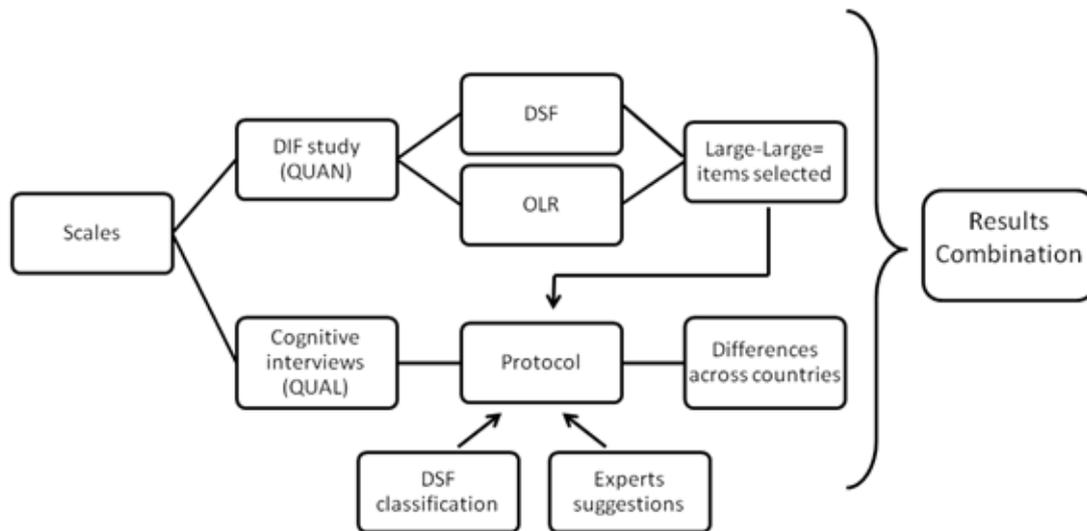
analysts were clear and by referring to the fourth party, the adjudicator, to take the final decision when they were not so clear.

Once the interviews were coded and checked, we analyzed the interpretations made by participants of the different groups. To do this, we compared the coding of themes and subthemes drawn from the narratives of participants in each group in each of the items selected in the first phase of the study.

Mixed Research framework design

The characteristics of this design respond to what Creswell (1995) classified as a study "QUAN + QUAL". The nomenclature reflects that there are two sequential phases, the first to run being the quantitative, but both equally relevant to the objectives of the study. Figure 1 shows the design characteristics graphically.

Figure 1. Phases of the (QUAN+QUAL) study.



Results

Quantitative findings

After performing the DIF analyzes separately for each method, the convergence between the results was examined. Convergence was evaluated by considering both methods and effect size measures results when evaluating overall DIF. For the next steps of the study, only items flagged with large DIF in both comparisons were considered in order to reduce uncertainty in the DIF empirical results. Table 1 includes the characteristics of items flagged: the item codes (composed by an acronym from the scale number and the number in the scale), the item contents and the scale to which they belong.

Table 1. Items with Large DIF.

Item code	Item content	Scale
Enj5	<i>"I am interested in learning about broad science"</i>	Enjoyment of science
Eff8	<i>"Identify the better of two explanations for the formation of acid rain"</i>	Science self-efficacy
Gen1	<i>"Advances in broad science and technology usually improve people's living conditions"</i>	General value of science
Gen3	<i>"Advances in broad science and technology usually help improve the economy"</i>	General value of science
Gen5	<i>"Advances in broad science and technology usually bring social benefits"</i>	General value of science
Per5	<i>"When I leave school there will be many opportunities for me to use broad science"</i>	Personal value of science
Ins1	<i>"Making an effort in my school science subjects is worth it because this will help me in the work I want to do later on"</i>	Instrumental motivation to learn science
Con4	<i>"School science topics are easy for me"</i>	Science self-concept

As shown in Table 1, eight items were classified as having Large DIF by both methods and in both comparisons. DSF analyzes were performed on these eight items in order to obtain information on the nature and location of the DIF effects. Table 2 shows the DSF form for each item determined from the criteria of Penfield, et al. (2009a), which considers the number of steps with DIF, the magnitude and sign. It also specifies the elements of the item to be reviewed in light of the classification carried out. The table shows a summary of the results found in the two comparisons used (Spain 1 - US 1 and Spain 2 - US 2) noting (*) the points where significant discrepancies were detected between the two.

Table 2. Summary of DSF form results.

Items	Steps	Magn.	Sign	Form	Revision
Enj5* ¹	1	Medium	US	Pervasive - Convergent	Item Level
	2,3	Large			
Eff8	1,2, 3	Large	US	Pervasive - Constant	Item Level
Gen1* ²	1,2,3	Large	US	Pervasive - Constant	Item Level
Gen3* ¹	1	Medium	Spain	Pervasive - Divergent	Item Level
	2,3	Large	US		
Gen5	1,2	Large	US	Pervasive - Convergent	Item Level
	3	Medium			
Per5	3	Large	Spain	Non- Pervasive - Constant	Score Level
Ins1	2	Medium	Spain	Non- Pervasive - Constant	Score Level
Con4* ³	1,2	Large	US	Pervasive - Convergent	Score Level
	3	Medium			

*1 Enj5 and Gen3 were classified in the comparison Spain 1 - US 1 as Non-Pervasive because there was no DIF detected in step 1. *2 Gen1 had a convergent constancy in the comparison Spain 2 - US 2 because the magnitude of DIF in steps 2 and 3 was medium. *3 Con4 was classified in comparison Spain 2 - US 2 as Non-Pervasive because there was no DIF detected in step 3.

Six items (Enj 5, Eff8, Gen1, Gen 3, Gen5 and Con 4) are pervasive as all steps displayed a substantial DSF against the US samples across comparisons. The sign of DSF effect for these items means that the advance from the lowest score levels to a higher score level is easier for the Spanish respondents than for the US respondents (except for Gen3 in step 1). The DSF of two of these items (Eff8 and Gen1) was constant, as the sign and magnitude were equal across steps. Items Enj5, Gen5 and Con4 showed a convergent DSF since the sign of the DSF was equal but magnitude changed across steps; while for item Gen1 the constancy of DSF varied across comparisons, being constant in one of them and convergent in the other one. Finally, in item Gen3, constancy was classified as divergent because the sign was different across steps. Following the advice by Penfield, et al. (2009a), context experts should look at the item stem to identify DIF causes because the DIF effect is located at the item level.

In the two items in which the DIF was classified as Non-pervasive, it was also constant (Per 5 and INS1). Non-pervasive DIF points out that the DIF effect could be located at the score levels. In item Per 5 and Ins1 the disadvantaged group was Spain which indicates that the transition from the lowest score levels to a higher score level was relatively easier for the US participants than for the Spanish participants. Content experts should review score levels to identify DIF causes.

Qualitative finding

The data analysis focused on the comparison between the narratives of the participants of the different groups. Firstly, it was noted that three of the items with large DIF (Gen1, Gen3 and Gen5) belonged to the same scale "General value of science". The wording of these three items contain a common element, the expression "Advances in broad science and technology." Reviewing the transcripts and the themes coded to the description of the meaning of this expression in each of the groups, we observed that the Spanish participants often referred to daily aspects such as mobile phones or the Internet, while Americans focused on broader issues and "new things that affect people" such

as development of new energy and new security systems. In addition, the Spanish participants more often used the subtheme "inventions" to define the term "Advances" and the Americans referred to "evolution and improvement". Table 4 shows transcripts that illustrate these differences.

Table 4. Finding about differences in the interpretation of "Advances in broad science and technology"

Participant codes	Narrative	Finding
Spain 4	<i>"When a new invention is done, daily life is easier... for example the light bulb or the television"</i>	Specific (daily life things) & inventions
Spain 7	<i>"In medicine advances, in technology, in machines which make easier our life, for example the refrigerator or other electrical appliances"</i>	Specific (daily life things) & inventions
US 5	<i>"It normally does help people's living conditions because just think like solar powered energy, like tablets and things like that that you can use to help your living, cut down on all the wasted energy we use and it normally helps people living conditions"</i>	General (global things) & evolution or improvement
US 13	<i>"Advances mean like a new finding or something good or bad that they have found in a study that could affect people. Again, like sickness or anything maybe they found a good advancement that this combination of drugs helps out more, or they have a bad advancement saying it doesn't work at all and they finally have the details to back something up"</i>	General (global things) & evolution or improvement

In addition to these differences, it is also noted that in item Gen5 the participants interpreted the term "social benefits" differently. The narratives of this item show how the Spanish were the only participants who referred to health services as a social benefit. These differences may reflect that the term has different meanings among participants of the different groups because of differences in each country's health and social services. Table 5 shows some examples of the different interpretations of the participants in the two groups.

Table 5. Examples of narratives in item Gen 5.

Participant codes	Narrative
Spain 21	<i>"For example, there are a lot of sick people and I have thought... advances in science are related to finding treatments or more information for improving their quality of life"</i>
Spain 1	<i>"... in everything... how the society helps me to have a better quality of life... for example, the medicine help people...."</i>
Spain 2	<i>"For example I have thought of hospitals because thank to science more lives are saved and that is a social benefit"</i>
US 3	<i>"like cars, so they have cars and a social benefit of having a car is you could get to more places you could help people get to places and it improves the transportation rate but then some people really don't have any cars"</i>
US 8	<i>"Well advances in technology like cell phones and computers and ways to communicate with other people even if they're really far away like across the world, you're still able to communicate with them because of that technology and that's a social benefit."</i>
US 1	<i>"Well, the first thing that I thought of with Science and Technology, social the first thing I thought of was Facebook. Bringing people back together who haven't seen or spoken to each other since college or something like that"</i>

There were also differences in item Per5, in which participants from different groups differently interpreted the phrase "When I leave school". The narratives showed that in both groups references to "career" were done as a theme, but Spanish participants developed topics closer to them in time, like future experiences at university, increasing the choice of the subtheme "university"; while American students spoke about later circumstances, such as situations at work, using the sub-theme "job" more frequently. Table 6 shows examples of this situation.

Table 6. Examples of the “university-job” sub-theme

Participant codes	Narrative	Sub-theme
Spain 5	<i>“When I am older I will study a science related career”</i>	University
US 11	<i>“When I have a job, I would be using science to be able to do that”</i>	Job
Spain 7	<i>“I have though that when I finish high school I won’t use science in my university studies”</i>	University
US 12	<i>“As I said, I probably will become a CSI investigator so, I was thinking of that”</i>	Job

Investigating the different temporal references used by participants from different groups, a new finding was discovered. Although both groups used the term “career”, due to the meaning of this term in Spanish, participants were making different interpretations. In Spanish, the term “career” is used for “professional career” in the same context as the English word, but also for referring to “studies at university”. In Spain, to enroll in universities is called in a general way “to do a career”. If Spanish participants are thinking of this second meaning, the situations narrated will be related to a different moment in time from the American participants. In addition, if Spanish participants are thinking about the University studies when responding, they can be thinking of different applications of the “broad science” expression from American participants who were thinking of the use of science in labor situations. It is also observed in American transcripts like US12 in which the participant declared *“As I said, I probably will become a CSI investigator so, I was thinking of that”*. Usually, science is present in university subjects but it is less frequently present in labor situations, which could explain the differences in the categories chosen across groups.

On analyzing the results of the CI, in addition to information concerning the comparison of the groups, some data on the influence of sociodemographic

characteristics in all participants were also collected. For example, we observed that the older participants, both from U.S. and from Spain, made more reference to college and future academic situations than younger participants. It was also found that children of parents with higher education obtained higher scores on the scale of "science self-efficacy" in both groups.

Findings integration

The integration of the results aims to connect the evidence obtained with analysis of DIF and the CI. The indications of Penfield, et al. (2009a) on the DSF form were followed on the reviewed items to combine the results of both methods.

Firstly, we compared the evidence on the items that had been classified as Pervasive DIF (Enj 5, Eff8, Gen1, Gen 3, Gen5 and with 4), focusing the review on the content of the item as directed by Penfield, et al. (2009a). Looking at the items with these characteristics it was found that three of them (Gen1, Gen 3 and Gen5) contained the phrase "Advances in broad science and technology" that the CI had identified as problematic because it generated different discourses from the participants in each group. In these items the participants of the various groups referred to different themes with the US being the disadvantaged group. Furthermore, in item Gen5 another expression, "social benefits" was located, that participants of different groups interpreted differently. The CI showed how participants gave different meanings to this term which was also present in the item content. Therefore, the identified DIF noted aspects of the content that, as well as being identified by experts during the expert appraisal, showed differences in the response processes made by the participants of different groups.

Moreover, in the items where the DSF was classified as Non-pervasive (Per5 and INS1) it was necessary to address the response categories (Penfield, et al. 2009a). The connection between the quantitative and qualitative results was observed for example in item Per5 in which, as shown in Table 3, the DIF was located in the third step prejudicing Spanish participants and with a Large

magnitude. In this item, the CI showed that participants were using a different time scheme that made them interpret the contents of the item differently, causing the Spanish participants to think of situations closer related to the short-term academic context, while American participants referred to situations in the longer term future related to work situations. The differences in the use of categories may be related to the fact that the academic context may provide more opportunities to use science, which is present, one way or another, in most specialties, therefore, if the Spanish are thinking in this context their answers will be less extreme, ie it is more difficult for them to answer Strongly Disagree. However, if the American participants are thinking in the work context they would like to enter, they can clearly determine whether they will use science or not, and therefore are more able to use the extreme category. An example is the US17 narrative when participant said: *"I really only think of myself using science in school, so when I get out, like I'm not going to do a job with science"*.

Also some of the items were ranked differently in the two comparisons, so their revision affects both the content of the item and the categories. For example, in item Enj5 the phrase "I am interested" led to differences in the interpretations of the participants because in Spanish it has two meanings. The differences of the participants were due to the Spanish participants not interpreting the expression in the shared sense, but instead thought about the other meaning of the term that qualifies something as advisable. The different interpretation of this term can be related both to the content of the item and the use of the categories. That is, it is necessary to review the item because it contains an element that generates different interpretations, an element that could be avoided by using a term that is equivalent in both versions.

However, it also affects the use of the categories, since the Spanish participants are thinking about something that suits them and that convenience depends on their plans for the future, ie they can clearly determine whether or not it suits them based on their experiences. Nevertheless, the interest in the

meaning given by the American participants refers to something more stable. Science may be of interest regardless of the use made of it. Therefore, the American participants may have more difficulties moving to Disagree and Strongly Disagree categories (since DSF data show DIF in steps two and three). The combination of quantitative and qualitative data shows, in this case, how the presence of a term that generates different interpretations may affect the interpretation made by the participants of different groups and the use of response categories. The evidence shows that participants from different groups are thinking about different aspects of science and are using a different scheme based on global aspects in the case of Americans and more specific aspects in the case of the Spanish.

Discussion

The study aimed to propose an approach for the analysis of the causes of DIF in the context of MR and examine its benefits over other approaches, a goal that was motivated by the inadequacy of the methodologies used so far. To do this, we designed the research to include a quantitative part targeted at the detection of DIF and a qualitative part to investigate the interpretations made by the participants. The combination of the results of both phases has focused on finding the items that generate the differences between the question-and-answer processes of the participants in each national group, in order to make inferences about the possible causes of DIF. The novelty of this study is to apply an MR design that combines statistical methods and CI to study the causes of DIF. The combination of both methods has shown the possibility of using CI to interpret the results from the DIF, and the connection between the quantitative and qualitative results.

As reflected in the results, the MR has proven to be a suitable paradigm for studying the causes of DIF, as well as offering a more integrated view of the phenomenon, it has led to conclusions that could not be obtained by applying a single method. Furthermore, the Mixed Method design used for this study has facilitated the integration of the results provided by both methods as they

progressed through the phases. That is, each of the findings of the quantitative phase has focused the analysis the qualitative phase. In turn, each of the results provided by the CI was able to connect to the quantitative results for the items with DIF. The results reflect a connection between the characteristics of DIF and the participants' narratives, so that different types of DSF showed different patterns of interpretations in groups. Specifically it was observed how the items with pervasive DIF led to different response processes to the elements included in the text of the item, ie, people thought about different issues when they responded to these items. However, when the items were classified as non-pervasive DIF, participants considered different situations and thus the transition from one category to the next required different skill levels for each of the groups. Therefore, one can say that the items with pervasive DIF contain different conceptual elements that carry different meanings, while the items with non-pervasive DIF responses are different because of different experiences related to the context or culture. Results also suggest the presence of possible problems in the adaptation that are caused by the use of terms that are not equivalent in the two versions.

As for the contribution of each phase, in the DIF study, the use of two different methods to analyze the DIF as well as the two comparisons between the groups and the implementation of two effect size measures provided greater confidence in the results, which is especially relevant as the next phases of the study were based on these results. Furthermore, the combination of procedures contributed to, as defined in the context of MR, the weaknesses of one of the methods being compensated by the strengths of the other, thereby obtaining more stringent data (Bryman, 1988). Another justification for the application of two methods is that empirical studies that analyze DIF usually recommended using more than one procedure to increase confidence in the results, which may be threatened by the inherent limitations of statistical methods (Hambleton, 2006, Hidalgo, & Gomez-Benito, 2010).

In turn, CI have provided information which would not have been accessed if the statistical methods for DIF detection had been implemented separately. In addition, the involvement of the respondents means an advance over traditional methods of judgment. In this study, the experts made it possible to focus the investigation of interviews on the relevant points, but the information provided was not as complete as that obtained by the CI. The CI have given applied evidence on how the problems identified by experts manifest.

In relation to the evidence obtained on the equivalence of the versions, both the study of DIF and the CI have shown aspects that threaten the comparability of the groups. Furthermore, the combination of the results showed the scale "General Value of Science" as problematic, identifying three of its items in which differences in the response processes of the participants of different groups were observed with the expression, "Advances in broad science and technology"; and a large magnitude of DIF. Consequently, inferences about the differences or similarities between groups in the items of this scale should be established with caution.

The main difficulty of this study in terms of content has been the connection between the results of DIF and the evidence provided by the CI. The key to this difficulty lies in that when analyzing attitudinal scales, the interpretation of the use made of the response categories is much more complex than in the context of aptitude or performance tests, where the response alternatives of polytomous items may be different, while attitudinal scales always use the same labels, following the Likert model.

Future research will be directed at placing the methodological findings in relation to the usefulness of CI for locating the causes of DIF. In addition, further investigations will focus on locating the causes of DIF associated with less powerful effect sizes in order to narrow down all the possible sources, as despite having a high level of confidence in the results, other conditions should be explored to identify other causes and relate them to other levels of DIF.

In conclusion, the fundamental contribution of this work is the illustration of a new way of addressing the search for the causes of DIF within the MR paradigm, which involves the people who respond. Despite these limitations, this study shows a new approach to the study of the causes of DIF that is methodologically rigorous and which has established some conclusions relevant to the advancement of the field.

References

- Allalouf, A., Hambleton, R.K., & Sireci, S. (1999). Identifying the Causes of DIF in Translated Verbal Items. *Journal of Educational Measurement*, 36 (3), 185-198.
- Beatty, P., & Willis, G.B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71 (2), 287-311.
- Bryman, A. (1988). *Quantity and Quality in social research*. London, Boston: Unwin Hyman publications.
- Clarke, P. N., & Yaros, P. S. (1988). Research blenders: Commentary and response. *Nursing Science Quarterly*, 1, 147-149.
- Creswell, J. W. (1995). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Edmeades, J., Nyblade, L., Malhotra, A., MacQuarrie, K., Parasuraman, S., & Walia, S. (2010). Methodological Innovation in Studying Abortion in Developing Countries: A “Narrative” Quantitative Survey in Madhya Pradesh, India. *Journal of Mixed Methods Research*, 4 (3), 176-198.
- Elosua, P., & López-Jauregui, A. (2007). Potential Sources of Differential Item Functioning in the Adaptation of Tests. *International Journal of Testing*, 7 (1), 39-52.
- Ercikan, K.; Arim, R.; Law, D.; Domene, J. Gagnon, F., & Lacroix, S. (2010). Application of Think Aloud Protocols for Examining and Confirming Sources of Differential Item Functioning Identified by Expert Reviews. *Educational Measurement: Issues and Practice*, 29 (2), 24-35.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4, 113-148.

- Gierl, M.J., & Khaliq, S.N. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement*, 38 (2), 164-187.
- Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care*, 44 (11), S182-S188.
- Hidalgo, M. D., & Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (3rd edition)*. USA: Elsevier - Science & Technology.
- International Association for the Evaluation of Educational Achievement- IEA (2011). Progress in International Reading Literacy Study (PIRLS).
- International Test Commission (2010). International Test Commission Guidelines for Translating and Adapting Tests. Retrieved January 13, 2012, from <http://www.intestcom.org>.
- Linn, R. L., & Harnisch, D. L. (1981) Interactions between item content and group measurement on achievement test items. *Journal of Educational Measurement*, 18, 109-118.
- Miller, K. (2007). *Design and Analysis of Cognitive Interviews for Cross-National Testing*. European Survey Research Association Annual Meeting. Prague, Czechoslovakia.
- Miller, T.R., & Spray, J.A. (1993). Logistic Discriminant Function Analysis for DIF Identification of polytomously scored items. *Journal of Educational Measurement* 30 (2), 107-122.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 17, 297-334.
- National Center for Health Statistics (NCHS). Q-notes software. Available in <https://wwwn.cdc.gov/qnotes/login.aspx>.

- Organisation for Economic Co-operation and Development. (2011). Programme for the International Assessment of Adult Competencies (PIAAC).
- Organisation for Economic Co-operation and Development. (2006). PISA 2006 database. Downloaded from the <http://pisa2006.acer.edu.au/downloads.php> on July 20, 2010.
- Organisation for Economic Co-operation and Development. (2009). Program for International Student Assessment (PISA).
- Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*, 150-151.
- Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items. *Journal of Educational Measurement, 47*, 129-149.
- Penfield, R.D., Alvarez, K., & Lee, O. (2009a): Using a Taxonomy of Differential Step Functioning to Improve the Interpretation of DIF in Polytomous Items: An Illustration. *Applied Measurement in Education, 22 (1)*, 61-78
- Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009b), An NCME Instructional Module on Using Differential Step Functioning to Refine the Analysis of DIF in Polytomous Items. *Educational Measurement: Issues and Practice, 28*, 38-49.
- Presser, S., Rothgeb, J., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., & Singer E. (2004). *Methods for Testing and Evaluating Survey Questions* (pp. 45-66). NJ: John Wiley & Sons.
- Roussos, L.A., & Stout, W.F. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-71.
- Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 64-96), Baltimore, Maryland: The Johns Hopkins University Press.

- Shealy, R., & Stout, W.F. (1993). An item response theory model for test bias. In P.W. Holland & H.Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice*, 16 (1), 12-19.
- Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R.K. Hambleton, P.F. Merenda y S.D. Spielberger (eds.): *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). New Jersey: Lawrence Erlbaum Associates.
- SPSS, Inc. 2007. SPSS-16 User's guide. Chicago, USA.
- Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27, 53-75.
- Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Van de Vijver, F.J.R., & Poortinga, Y.H. (2005). Conceptual and methodological issues in adapting tests. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Hillsdale, NJ: Lawrence Erlbaum.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.
- Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.). *Differential Item Functioning* (pp. 337-347). Hillsdale. NJ: Erlbaum.
- Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for*

Binary and Likert-Type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In R.W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). IAP - Information Age Publishing, Inc.: Charlotte, NC.

DISCUSIÓN

El objetivo general de la tesis fue plantear la evaluación de la calidad de las mediciones aportadas por escalas y cuestionarios, mediante diseños de investigación mixtos en los que los métodos de pretest cognitivo y los métodos psicométricos se combinaran con el fin de obtener evidencias de validez. Para ello, el primer paso fue enmarcar dicha evaluación en los contextos de la investigación mediante encuestas y de la evaluación psicológica. Una vez situados en estos contextos, la evaluación de la calidad se planteó como parte de un proceso de obtención de evidencias de validez basadas en los procesos de respuesta de los participantes. Las evidencias de validez basadas en los procesos de respuesta fueron obtenidas mediante métodos de pretest cognitivo, cuyos resultados fueron combinados con los resultados proporcionados por los métodos cuantitativos utilizados. Con el propósito de guiar la combinación de los datos y de obtener conclusiones más completas, el diseño de cada uno de los estudios se realizó siguiendo los fundamentos de la investigación mixta.

Para alcanzar el objetivo general se plantearon objetivos específicos que fueron abordándose en los distintos estudios. El diseño de los estudios permitió obtener información sobre tres cuestiones fundamentales:

a) La utilidad de los métodos de pretest cognitivo para solucionar problemas novedosos, como pueden ser la evaluación de la convergencia entre distintos tipos de informantes en el caso de la Codificación del Comportamiento (CC), o la interpretación de resultados psicométricos en el caso de las Entrevistas Cognitivas (EC).

b) La capacidad de los métodos de pretest cognitivo para aportar evidencias de validez basadas en los procesos de respuesta. La importancia de este punto subyace en el hecho de que este tipo de evidencias ha sido incorporado en la última revisión de los *Standards*, y en que aún no se ha alcanzado un

consenso amplio sobre los métodos más adecuados para acceder a ellas.

c) Los beneficios derivados de la combinación de los resultados de los métodos de pretest cognitivo con los procedentes de otros métodos cuantitativos, especialmente métodos psicométricos, frente a su utilización en exclusividad. Los estudios realizados buscan demostrar las ventajas de utilizar varios métodos en comparación con la utilización de un método único.

Los resultados y conclusiones extraídas de los estudios realizados han permitido dar respuesta a los interrogantes planteados y proponer algunas conclusiones generales.

Utilidad de los métodos de pretest cognitivo para resolver problemas de investigación

En relación a la utilidad de los métodos de pretest cognitivo, los estudios 1 y 2 proporcionaron información sobre la CC, y los estudios 3 y 5 sobre las EC.

Sobre la CC, los Estudios 1 y 2 mostraron su utilidad para determinar la calidad de las respuestas emitidas por los informantes proxy y para evaluar la convergencia entre sus respuestas y las proporcionadas por los informantes directos. La información extraída mediante este método sobre la calidad de las respuestas de los proxies ha proporcionado elementos que llevan a dudar de algunas conclusiones previas recogidas en la literatura. Estudios anteriores plantean que existe una relación clara entre la falta de precisión de las respuestas de los proxies y el hecho de que éstos informantes no sean los informantes directos. La CC ha permitido relacionar esta falta de precisión con características del cuestionario presentes tanto en la estructura como en el contenido de las preguntas. Por ejemplo, el Estudio 1 mostró como respuestas

inadecuadas de los proxies surgen ante un mal diseño de las alternativas que induce a dar una respuesta incompleta, o ante un énfasis inadecuado de elementos irrelevantes de la pregunta.

En el Estudio 2, la contribución principal de la CC se basa en las mejoras obtenidas en el proceso de evaluación de la convergencia entre informantes, frente a los procedimientos utilizados habitualmente. La CC generó categorías específicas que permitieron recoger todos aquellos comportamientos de interés para el estudio. En concreto, se desarrollaron códigos para clasificar los distintos tipos de respuestas que dieron los informantes con el objetivo de conocer detalles relevantes sobre el tipo de desacuerdo, es decir, facilitó la indagación en los motivos de la no-convergencia. La aportación de la CC se obtuvo gracias a una de las características básicas de este método, su flexibilidad de adaptación a los objetivos planteados en el estudio. En este sentido, la flexibilidad de la CC ha abierto nuevas posibilidades en cuanto a la utilización de categorías específicas, lo que supone un avance frente a evaluaciones previas de la convergencia entre informantes. El Estudio 2 proporcionó un nuevo esquema para determinar las situaciones de acuerdo y desacuerdo entre informantes. Un esquema en el que es posible conocer las causas del desacuerdo y por tanto establecer conclusiones sobre las diferencias en los procesos de respuesta que realizan informantes directos y proxies.

Los Estudios 3, 4 y 5 trataron de mostrar la utilidad de las EC para completar e interpretar los resultados procedentes de análisis psicométricos y análisis del Funcionamiento Diferencial de los Ítems (DIF). En el Estudio 3, se relacionaron los resultados de análisis estadísticos y de la dimensionalidad de una escala psicológica con el tipo de narrativa desarrollada por los participantes. El propósito era observar la relación existente entre ambos tipos de datos. Los resultados mostraron que una alta variedad de temáticas presentes en los discursos de los entrevistados se relacionaba con una mayor heterogeneidad en las respuestas de los participantes. La variabilidad en las respuestas a la escala se manifestó a través de valores altos en los índices de

discriminación y evidencias más sólidas de unidimensionalidad (altos porcentajes de varianza explicada por el primer factor).

El Estudio 5 tuvo una estructura similar pero en este caso los resultados cuantitativos fueron obtenidos mediante el análisis del DIF. Una descripción detallada del proceso mediante el cual se analizó el DIF se presenta en el Estudio 4, donde se muestra cómo se aplicaron dos procedimientos y dos medidas del tamaño del efecto con el fin de tener una mayor seguridad en los resultados. En el Estudio 5 se aplicaron las EC para conocer las causas que habían provocado DIF en los ítems seleccionados en el Estudio 4. En este caso, el propósito era relacionar las narrativas obtenidas mediante las EC con la presencia de DIF. Se observó que aquellos ítems que funcionaban diferencialmente entre los grupos, daban lugar a discursos diferentes entre los participantes de dichos grupos. Estos resultados permitieron de nuevo relacionar las diferencias en los resultados cuantitativos con diferentes interpretaciones de los participantes, además de conectar esas diferencias en las interpretaciones con características específicas de los grupos. Los estudios 3, 4 y 5 han aportado información relevante para concluir que las EC son útiles para dar significado a los resultados psicométricos permitiendo una interpretación más completa de los mismos.

Capacidad de los métodos de pretest cognitivo para aportar evidencias de validez basadas en los procesos de respuesta

Los estudios incluidos en esta tesis muestran también la capacidad de los métodos de pretest cognitivo para extraer evidencias de validez sobre los procesos de respuesta que realizan los participantes.

Los estudios que utilizaron la CC proporcionaron información detallada sobre los comportamientos de los participantes durante la interacción entrevistador-entrevistado. Esta información facilitó el planteamiento de

inferencias sobre los posibles procesos de respuesta realizados por los informantes, ya que la codificación mostró los puntos problemáticos y las posibles interrupciones ocurridas durante el proceso. A su vez, los detalles sobre los procesos de respuesta realizados para responder informaron sobre el ajuste entre el constructo pretendido y la respuesta recogida, lo que dio lugar a la formulación de evidencias de validez.

Por su parte, los estudios que utilizaron EC mostraron resultados que reflejan claramente los procesos de respuesta realizados por las personas cuando responden a las preguntas de una escala o cuestionario. De ese proceso de respuesta, las EC extrajeron los indicadores del constructo (i.e., patrones de interpretación), presentes en las respuestas de los participantes, lo que permitió indagar sobre la procedencia de informaciones no esperadas. Por ejemplo, al estar la escala objetivo diseñada para evaluar la función familiar, se esperaba un mayor ajuste del constructo en personas que convivían con sus familiares en el mismo domicilio. Sin embargo, los resultados psicométricos mostraron mejor funcionamiento de los ítems en las personas que vivían solas. Esta contradicción pudo ser interpretada gracias a las evidencias obtenidas durante las EC. En las narrativas de los participantes se observó que las personas que vivían solas mantenían su concepto de familia inmóvil durante las respuestas a todas las preguntas, es decir, siempre pensaban en lo mismo. Sin embargo, aquellas personas que no vivían solas variaban su concepto de familia a lo largo de los ítems, es decir, en ocasiones hacían referencia a otros miembros del hogar y en ocasiones nombraban a familiares con los que no vivían e incluso amigos. Por tanto, los resultados de las EC han proporcionado evidencias sobre el ajuste entre la respuesta dada por los participantes y el constructo que se pretendía medir, es decir, han proporcionado evidencias basadas en los procesos de respuesta tal como se refleja en las fuentes de validez enumeradas en los *Standards*.

¿Cuáles son los beneficios de combinar resultados siguiendo los fundamentos de la investigación mixta?

En todos los estudios realizados se pueden observar, de una manera u otra, los beneficios de utilizar varios tipos de métodos o varios tipos de datos. Los Estudios de Modelos Mixtos, Estudios 1 y 2 que aplican la CC, muestran una alta interrelación entre los aspectos cualitativos y cuantitativos. La aproximación cualitativa se hace presente en las primeras fases del estudio: durante la definición de las categorías que se elaboran para reflejar los comportamientos ocurridos durante la interacción entrevistador-entrevistado; y durante la asignación de códigos a las secuencias ocurridas durante la administración. En ambos estudios, la aproximación cuantitativa se aplica durante el análisis de los datos. En esta fase, los “números” permiten resumir los resultados en forma de frecuencias y tablas de contingencia, herramientas que facilitan la localización de los aspectos problemáticos del instrumento, así como la comparación entre los ítems más y menos problemáticos.

El Estudio 4 nos ofrece una perspectiva diferente focalizada en las ventajas de utilizar distintos procedimientos dentro de un mismo paradigma. En este estudio, se compararon los resultados procedentes de dos procedimientos cuantitativos distintos. Las ventajas de combinar sus resultados son claras ya que es posible comparar las conclusiones obtenidas en este estudio con las que hubieran resultado si se hubiera aplicado uno sólo, cualquiera, de los dos procedimientos. Los datos muestran que el número final de ítems con DIF tras la combinación fue notablemente menor que el número de ítems con DIF obtenido por cada uno de los procedimientos. Por tanto, la aportación de la investigación mixta es, en este estudio, el incremento de la confianza en los resultados provocado por la evaluación de la convergencia entre los procedimientos cuantitativos aplicados.

Por otro lado, en los Estudios 3 y 5 se reflejan claramente los beneficios de la investigación mixta en cuanto a la posibilidad de alcanzar conclusiones más

completas combinando resultados. En ambos estudios las EC realizaron una gran aportación para la interpretación de resultados cuantitativos, procedentes de los análisis psicométricos habituales en el caso del Estudio 3 y del análisis del DIF en el Estudio 5. En el Estudio 3, los datos obtenidos durante las EC se centraron en relacionar los discursos desarrollados por los participantes con las respuestas que habían dado a la escala. Se observó que, las propiedades psicométricas de los ítems variaban en la misma dirección que las interpretaciones de los participantes, es decir, una mayor variedad de contenidos en los discursos se manifestaba en una mayor variedad de respuestas. En el Estudio 5, el análisis de las EC estuvo más focalizado en conectar la presencia de DIF en los ítems con discursos diferentes entre los distintos grupos. El objetivo fue detectar patrones de interpretación similares entre los participantes de un mismo grupo, y diferentes a los de los participantes del otro grupo. De esta forma, el análisis pormenorizado de las interpretaciones de los participantes permitió detectar los elementos que pueden estar asociados al DIF y por lo tanto, plantear hipótesis sobre sus posibles causas.

De forma general podemos decir, que la aportación de la investigación mixta es visible en todos los estudios de esta tesis, ya que gracias a ella se han podido conectar los resultados procedentes de los distintos métodos aplicados.

CONCLUSIONES Y LINEAS FUTURAS

Conclusiones finales

Para concluir, es importante señalar que tanto el análisis individual, de los resultados de cada estudio, como el global permiten afirmar que:

- a) Los métodos de pretest cognitivo son útiles para resolver problemas metodológicos más complejos que los que habitualmente se han abordado cuando se aplican de forma rutinaria durante el pretest de los cuestionarios de encuesta. Tanto la CC como las EC han mostrado, a lo largo de los estudios de esta tesis, una mayor amplitud de posibilidades en cuanto a los análisis que permiten y los resultados que proporcionan.
- b) Los métodos de pretest cognitivo aportan evidencias de validez valiosas sobre los procesos de respuesta de las personas que responden a los cuestionarios y escalas. Las evidencias proporcionadas por la CC han aportado datos sobre el ajuste entre el constructo evaluado y las respuestas de los participantes, y entre las respuestas de diferentes tipos de informantes. A su vez, las EC han proporcionado evidencias sobre la relación entre las respuestas a la escala y la interpretación de los participantes, y entre las características de los ítems y las interpretaciones de los participantes de distintos grupos en estudios comparativos.
- c) Utilizar distintos procedimientos dentro de un paradigma de investigación mixta facilita la obtención de conclusiones más ricas que las proporcionadas por los métodos de forma separada. Los estudios muestran diferencias entre los resultados obtenidos tras la combinación de métodos y los resultados que se hubieran obtenido con la utilización exclusiva de dichos métodos. La integración de resultados ha resultado beneficiosa en relación a varios aspectos como son: la cantidad de información obtenida; la

calidad de esa información, que es mucho más completa por contener elementos procedentes de distintos métodos; y la posibilidad de compensar las debilidades de un método con las aportaciones del otro. Esto último se refleja en los estudios cuando uno de los métodos mejora la interpretación de los resultados del otro o cuando la seguridad en los resultados se incrementa por haber sido corroborados con otro método.

Todo ello permite afirmar que la aproximación mixta es una buena alternativa para mejorar la evaluación de la calidad de las mediciones procedentes de escalas y cuestionarios; y que, dentro de esta aproximación mixta, los métodos de pretest cognitivo son útiles para obtener evidencias de validez basadas en los procesos de respuesta y para complementar la información proporcionada por otros métodos.

Limitaciones y líneas futuras

Al mirar atrás desde este punto, en que los estudios han sido finalizados y los resultados obtenidos, se es consciente de todos los pasos que se han dado, de las dificultades, de las limitaciones y de cómo se podrían plantear las próximas etapas. En todo el proceso de realización de la tesis, las mayores dificultades se han debido a la complejidad del objeto de estudio. El hecho de querer abordar un tema tan extenso, como la evaluación de la calidad de las mediciones desde una perspectiva metodológica amplia y novedosa, implicaba adoptar un esquema complejo como es el paradigma de investigación mixta.

La aproximación mixta ha proporcionado mucha riqueza sobre todo en la variedad y calidad de los datos a los que ha permitido acceder. Sin embargo, los diseños utilizados en los estudios han resultado costosos a todos los niveles. Como ya señalaban Johnson y Christensen (2008), las principales desventajas de

los diseños mixtos son que requieren mucho tiempo, recursos económicos y la presencia de varios investigadores, de manera que puedan desarrollarse varias etapas simultáneamente. Además, Johnson y Christensen (2008) también señalaron otro punto importante, la necesidad de dominar los métodos que se están aplicando y conocer sus fundamentos, de forma que los datos puedan combinarse adecuadamente. En el contexto en que se han realizado los estudios de esta tesis también han surgido estas dificultades. En muchos casos las carencias han podido solventarse realizando un mayor esfuerzo en cuanto al nivel de dedicación, pero en otros casos las actividades realizadas han tenido que adaptarse a los recursos disponibles, teniendo que renunciar, por ejemplo, a una mayor cantidad de información. Es el caso del Estudio 5, en el que el número de participantes fue limitado debido a los costes económicos que ocasionó la aplicación de EC en Estados Unidos.

Revisando individualmente los estudios, se puede intuir otra limitación. Al tratarse de una tesis con objetivos fundamentalmente metodológicos, el contenido “sustantivo” de los estudios ha sido seleccionado siguiendo criterios prácticos, es decir, era importante disponer, por ejemplo, de una escala que tuviera todos los elementos que queríamos analizar, pero no era especialmente importante el constructo que esa escala midiera. Este es el motivo por el que el contenido de los estudios no responde a una planificación previa, ya que en cada momento se ha elegido el instrumento más adecuado para poder investigar el fenómeno metodológico que nos interesaba. Por ejemplo, en los Estudios 1 y 2, se utilizaron datos procedentes de una encuesta de discapacidad cuyo pretest formó parte de un contrato I+D financiado por el Instituto Nacional de Estadística en España (INE; Ministerio de Sanidad y Consumo, 2007). Esta situación plantea ventajas como la posibilidad de realizar una investigación aplicada en las que se utilizó la metodología para responder a necesidad “real”, además de poder acceder a las personas que fueron reclutadas para participar en dicho pretest. Sin embargo, la desventaja principal es que el objetivo del trabajo del contrato no era resolver una cuestión metodológica, sino una de carácter aplicado: evaluar y optimizar el diseño de las preguntas del

cuestionario; lo que implicó planificar el estudio de forma que se pudiera aprovechar la experiencia aplicada para avanzar en la investigación metodológica. En el Estudio 3 ocurrió algo similar con la escala APGAR evaluada. De nuevo, este trabajo se realizó por un contrato del INE, que nos solicitó la realización del pretest de la Encuesta Nacional de Salud (Ministerio de Sanidad y Consumo, 2007), donde estaba inserta dicha escala. Por último, los Estudios 4 y 5 contaron con datos del Estudio PISA (Program for International Student Assessment; OECD 2009b), en concreto con algunas de las escalas incluidas en el Cuestionario del Estudiante. La selección de las escalas se realizó pensando en la amplitud de la información disponible y en la posibilidad de contar con instrumentos en diferentes idiomas. A pesar de las limitaciones que supone el tener que ajustar la investigación metodológica a las investigaciones aplicadas que hemos ido realizando, esta situación también contiene aspectos positivos. Y es que hemos podido contribuir a que se haga una mejor utilización de los instrumentos, tanto en el marco de los contratos donde los responsables de los estudios han incorporado la mayor parte de las sugerencias que hemos propuesto, como en el contexto del Estudio PISA, donde nuestros resultados pueden llevar a un uso más completo de las escalas evaluadas.

Por otra parte, a medida que se ha profundizado en la utilización de los distintos métodos, también han ido surgiendo nuevas inquietudes y necesidades que se plantearán en estudios futuros. Por ejemplo, las conclusiones obtenidas de los Estudios 1 y 2 sobre la versatilidad de la CC, han motivado que se plantee la utilización de este método en otros contextos, como por ejemplo para codificar comportamientos no verbales en el ámbito sanitario. El hecho de que la CC pueda ser adaptada a las necesidades del investigador y las categorías de codificación puedan ser desarrolladas para un objetivo específico, ha despertado el interés por este método en profesionales que quieren observar si los pacientes siguen las instrucciones que les proporcionan. Específicamente, el proyecto planteado consiste en codificar el comportamiento de pacientes diabéticos que han sido instruidos para auto-administrarse un

tratamiento, con el fin de detectar los puntos problemáticos y poder así mejorar los talleres dedicados a la enseñanza de esa auto-administración.

También la profundización en el estudio de las causas del DIF ha provocado que surjan nuevas propuestas en este contexto. La primera idea es realizar un estudio que permita confirmar las conclusiones establecidas en el Estudio 5. Se busca, por un lado, asentar los hallazgos metodológicos en relación a la utilidad de las EC para obtener evidencias de validez basadas en procesos de respuesta; y por otro, replicar los tipos de causas que se han desarrollado en este estudio mediante la aplicación de la misma estrategia de validación tanto a las administraciones del PISA más recientes como a las de otras posibles escalas.

Actualmente, se están planteando más estudios cuyo objetivo es continuar las líneas de investigación desarrolladas en esta tesis, que son cada vez más sólidas, y su avance supone cada día nuevos retos.

DISCUSSION

The overall objective of the thesis was to present an evaluation of the quality of measurements provided by scales and surveys, using Mixed Research (MR) design in which cognitive pretest methods as well as psychometric methods were combined to provide validity evidence. The first step was to frame the assessment in the context of survey research and psychological assessment. Once situated in this context, the quality assessment was proposed as part of a process to obtain validity evidence based on respondents' response processes. The validity evidences based on respondents' response processes, were obtained through cognitive pretest methods whose results were combined with results provided by quantitative methods used. In order to guide the combination of data and draw thorough conclusions, the design of each study was carried out following the principles of MR.

To achieve the overall objective, specific objectives were raised in the different studies. The design of the studies gave information on three key issues:

a) The use of cognitive pretest methods to resolve new problems, such as the assessment of the convergence of different types of informants in the case of Behavioral Coding (BC) or interpreting psychometric results in the case of Cognitive Interviews (CI)

b) The capacity of the cognitive pretest method to contribute evidence validity based on respondent's response processes. The importance of this point is the fact that this type of evidence has been incorporated in the final review of the *Standards*, and that an agreed method to access it is still without a broad consensus.

c) The benefits derived from the combination of findings from the cognitive pretest method along with the data drawn from other quantitative methods, especially psychometric, compared with their exclusive use. The studies carried out show advantages of using various different methods compared with using only one.

The results and conclusions extracted from the studies have been able to provide answers to the questions raised and some general conclusions suggested.

Use of cognitive pretest methods to resolve research problems

In relation to the use of cognitive pretest methods, Studies 1 & 2 provide information on BC, and Studies 3 & 5 on CI. Regarding BC, Studies 1 & 2 are useful in determining the quality of answers given by proxy reports and in evaluating the convergence of these answers with those provided by self-reporter. The information extracted with this method on the quality of proxy answers provided elements of doubt about conclusions previously reached in literature. Previous studies set a clear relationship between the lack of precise proxy answers and the fact that the reports are not direct. BC relates this lack of precision with the structure of the survey as much as with the contents of the questions. For example, Study 1 shows how inadequate proxy answers come about because of bad design in the possible alternatives, resulting in incomplete answers, or from placing inadequate emphasis on irrelevant elements of the question.

In Study 2, the main contribution of BC is based on the improvements made in the evaluation process of the convergence of informants compared with procedures normally used. BC generates specific categories that gather behaviors of interest to the study. Specifically, codes were developed to classify the different types of answers the informants gave with the objective of identifying relevant details about the type of disagreements, namely, it facilitated the inquiry into reasons of non-convergence. The BC contribution was thanks to one of the basic characteristics of this method: its flexibility and adaptation to the objectives set out in the study. In this sense, the flexibility of BC has opened up new possibilities when specific categories are used. This

represents an improvement on previous assessments on agreement or disagreement between informants. Study 2 provides a new scheme to determine agreement and disagreement situations among informants. This scheme makes it possible to identify the causes of disagreement and establish conclusions on the differences in the answering process of both self-reporter and proxy informants.

Studies 3, 4 and 5 deal with highlighting the use of CI to complete and interpret results drawn from psychometric Differential Item Functioning (DIF) analysis. Study 3 links the statistical analysis results and the dimensionality of a psychological scale with a type of narrative developed by the participants. The purpose was to observe the link between both types of data. The results show that a large variety of themes present in the interviewees discourses were related to a greater heterogeneity in the participants responses. The variety of responses to the scale was indicated by high values in the discrimination indexes and more solid uni-dimensional evidence (high percentages of variance explained by the first factor).

Study 5 is structured similarly but in this case the quantitative results were obtained from DIF analysis. A detailed description of the process by which the DIF was analysed appears in Study 4, showing how two methods and two measures of the size of the effect were applied in order of increasing confidence of the results. CI were carried out in Study 5 to find the cause of DIF in selected items of Study 4. In this case the purpose was to link the narratives obtained through CI where DIF is present. It was observed that these items worked differentially between the groups, giving way to differing discourses among participants of said groups. These results again show the differences in the quantitative results interpreted differently by the participants, furthermore, it connects these different interpretations with specific characteristics of the groups. Studies 3,4 and 5 contributed relevant information to conclude that CI are useful in giving meaning to the psychometric results, providing a fuller understanding of them.

Capacity of the cognitive assessment methods to provide validity evidence based on the response process

The studies included in this thesis also demonstrate the capacity of cognitive pretest methods in extracting validity evidence based on the respondents' response processes.

Studies that use CC provide detailed information on the participants behavior during interviewer–respondent interaction. This information facilitates the inference approach to possible response processes carried out by the informants because the coding highlights the problem points and possible interruptions that occurred during the process. On the other hand, details on the response process made when responding provide information on the fit between the intended construct and responses collected, which led to the formulation of validity evidence.

For their part, studies that used CI show results that clearly reflect the response process made by people when they respond to scale or survey questions. In this response process, CI extract the construction indicators (i.e. interpretation patterns), present in the participants responses, so the source of unexpected information could be investigated. For example, as the target scale was designed to assess family function, a better fit to the construct was expected in people who live with their families. However, the psychometric results show better performance on items in people that live alone. This contradiction could be interpreted by the evidence obtained during CI. In the participants narratives, it is noted that people living alone still maintained their family concepts in all their answers, and always thought the same thing. However, the family concept varies throughout the items among people that don't live alone, i.e., they sometimes make reference to other household members and sometimes to family members that they don't live with, and even to friends. Therefore, the results of the CI have provided evidence on the fit between the response given by the participants and the construct being

measured. In other words, it provided evidence based on the response process as reflected in the sources of validity evidence listed in the *Standards*.

What are the benefits of combining results following the guidelines of Mixed Research?

In all the studies carried out the benefits can be observed, in one way or another, of using different methods or different types of data. Mixed Models Studies, studies 1 and 2 that apply BC, show a high relationship between the qualitative and quantitative sections. The qualitative approach is present in the early stages of the study: in the definition of the categories that are made to reflect behavior during the interviewer–respondent interaction. It is also present in the assignment of codes to sequences that occur during administration. In both studies, the quantitative approach is applied for data analysis. In this stage, the “numbers” permit summarizing results in frequency and cross-tables, tools that facilitate the location of problematic aspects of the instrument, and the comparison between more and less problematic items.

Study 4 offers a different perspective focusing on the advantages of using different procedures with the same paradigm. In this study, the results from two different quantitative methods were compared. The advantages of combining their results are clear as it is possible to compare the findings obtained in this study with those results that would have been obtained had there been only one procedure applied, be it whichever one of the two. The data show that the final number of items with DIF, after being combined, was significantly less than the number of items with DIF obtained by each of the procedures. Therefore, the contribution of MR in this study is that it increased confidence in the results of the evaluation of the convergence of the quantitative procedures applied.

On the other hand, Studies 3 and 5 clearly reflect the benefits of MR regarding the possibility of achieving more complete conclusions by combining results. In both studies, CI made a large contribution to the interpretation of quantitative results from standard psychometric analyses in the case of Study 3 and the analysis of DIF in Study 5. Data in Study 3 obtained from CI focuses on linking the discourse developed by participants with the responses they made to the scale. It was noted that the psychometric properties of the items varied as much as the interpretation by the participants did i.e., a larger variety in the contents of the discourses was indicated by a larger variety in the responses. The CI analysis in Study 5 focuses more on connecting the presence of DIF in items with different discourses among different groups. The aim was to detect similar interpretation patterns among participants of the same group but different to the participants of other groups. Thus, the detailed analysis of participants' interpretations detected elements that could be associated with DIF and therefore suggest hypotheses for the possible causes.

In general we can say that the contribution of MR is visible in all the studies of this thesis, and therefore it has been possible to connect results from the different methods applied.

CONCLUSIONS AND FUTURE STUDIES

Final Conclusions

To conclude, it is important to note that both the individual analysis of the results of each study and the global analysis conclude that:

a) The cognitive pretest methods are useful for resolving more uncommon complex methodological problems that were found in routine assessment survey questionnaires. Both BC and CI show, throughout the studies in this thesis, a wider range of possibilities in both the analysis they allow and the results they provide.

b) Cognitive pretest methods provide valuable validity evidence based on respondents' response processes to questionnaires and scales. The evidence provided by BC contributed data on the fit between the evaluated construct and participants' responses, and also responses from different types of informants. In turn, CI provides evidence on the relationship between responses to the scales and the participants' interpretations, as well as between the characteristics of items and their interpretation by participants from different groups in comparative studies.

c) The use of different procedures within a MR paradigm allows firmer conclusions to be drawn than those provided by individual methods. The studies show differences between results from the combination of methods and the results that would have been obtained from those methods used alone. The integration of results have been beneficial in relation to various aspects such as: the amount of information obtained and the quality of that information, which is more complete because it contains elements of different methods as well as the weaknesses of some methods being offset with the contributions others. The latter is reflected in studies where one method improves the

interpretation of results from the other or when confidence in the results is increased because of the support of another method.

All of this suggests that the MR approach is a good alternative for improving the assessment quality of scales and questionnaire measurements. Also, that within this mixed approach, the cognitive pretest methods are useful for obtaining validity evidence based on the response process and for complementing information provided by other methods.

Limitations and future studies

In hindsight, with the studies completed and results obtained, one is aware of the difficulties and limitations in each step taken, and how the next steps could be set out. Throughout the process of writing the thesis, the main difficulties were due to the complexity of the subject matter. Tackling a subject as extensive as the assessment of the quality of measurements with a new and broad methodological perspective meant adapting a complex scheme, as the MR paradigm is.

The mixed approach has enriched the studies especially in the variety and quality of data accessible. However, the designs used in the studies proved costly at all levels. As already pointed out by Johnson and Christensen (2008), the main disadvantages of mixed designs is that they are very time consuming and require economic resources as well as several researchers for multiple stages to be developed simultaneously. Another important point from Johnson and Christensen (2008), is the need to master the methods that are being implemented and to know the basics so that data can be combined properly. In the context of how the studies in this thesis were carried out, these difficulties also arose. In many cases, the shortcomings were solved by increasing the level of dedication. In other cases the activities had to adapt to the available resources, for example, having to give up on obtaining a larger amount of

information. This is the case in Study 5, in which the number of participants was limited due to economic costs resulting in the application of CI in the United States.

Reviewing the individual studies, one can sense another limitation. As the objectives of this thesis are mainly methodological, the fundamental contents of the studies were selected following practical criteria, for example, it was important to have a scale that included all the elements we wanted to analyse, but the construct measured in this scale not especially important. This is why the contents of the studies do not correspond to prior planning, as the most appropriate tool was always chosen to research the methodological phenomenon that interested us. For example, we used data from a disability assessment survey in Studies 1 and 2 that was part of a R+D contract funded by the National Statistics Institute in Spain (NSI, Ministry of Health, 2007). This situation offers advantages such as the chance to do applied research in which methodology was used to meet real needs, as well as to gain access to the people recruited to participate in the aforementioned assessment. However, the main disadvantage is that the objective of the contracted work was not to resolve any methodological issue, but one of an applied character: to evaluate and optimize the design of the questionnaire. This involved planning the study so the experience could be used to advance applied research methodology. Something similar occurred in Study 3 with the evaluated APGAR scale. Again, this work was done under an NSI contract to complete the pretest of the National Health Survey (Ministry of Health, 2007), in which the scale was included. Lastly, studies 4 and 5 rely on PISA study data (Programme for International Student Assessment; OECD 2009b), specifically some of the scales included in the Student Questionnaire. The selection of the scales was done considering the depth of information available and the possibility of having instruments in different languages. Despite the constraints of having to adjust methodological research to the applied research we had been doing, it also has positive aspects. We have also been able to contribute to the better use of instruments, such as within the framework of contracts where the heads of

studies have incorporated most of the suggestions we proposed, and with the PISA study, where our results lead to a fuller use of the scales evaluated.

On the other hand, in the use of different methods, new concerns emerged that need to be addressed in future studies. For example, the findings from Studies 1 and 2 on the versatility of BC motivated us to think about the use of this method in other contexts, such as encoding non-verbal behavior in health studies. The fact that BC can be adapted to the needs of the researcher and the coding categories can be developed for a specific purpose, has raised interest in this method in professionals who want to see if patients follow the instructions they are provided. Specifically, the proposed project is to encode the behavior of diabetic patients who have been instructed to self-administer treatment, in order to detect problem points and to improve workshops dedicated to teaching that self-administration.

Also, the more profound study of the causes of DIF has brought about the emergence of new proposals. The first idea is to conduct a study that confirms the findings in Study 5. Firstly, it seeks to lay down the methodological findings regarding the use of CI to obtain validity evidence based on the response processes. Secondly, to replicate the types of causes developed in this study by applying the same strategy of validation to both the latest PISA administrations and those of other possible scales.

Currently, more studies are emerging whose objective is to continue the lines of research developed in this thesis, which are becoming more solid, and whose progress involves new challenges every day.

REFERENCIAS

REFERENCES

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Psychological Association, American Educational Research Association, y National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Beatty, P. C., y Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly*, 71, 287-311.

Brannen, J. (2005). *Mixed Methods Research: A discussion paper*. ESRC National Centre for Research Methods, Methods Review Paper.

Cannell, C. F., Fowler, F. J., y Marquis, K. H. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews. *Vital and Health Statistics*, 2 (26), 1-65.

Cho, Y. I., Martin, M. J., Conger, R. D., y Widaman, K. F. (2010). Differential Item Functioning on Antisocial Behavior Scale Items for Adolescents and Young Adults from Single-Parent and Two-Parent Families. *Journal of Psychopathology and Behavioral Assessment*, 32 (2), 157-168.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12, 229-238.

Creswell, J. W. (1995). *Research design: Qualitative and quantitative approaches*. Thousand Oaks. CA: Sage.

Ericsson, K. A., y Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215-251.

Eurostat Task Force. (2005). *Task force report on adult education survey*. Luxemburgo: Office for Official Publications of the European Communities.

- Foddy, W. (1996). *Constructing questions for interviews and questionnaires*. Cambridge: Cambridge University Press.
- Greene, J. C., Caracelli, V. J., y Graham, W. D. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, 11 (3), 255-274.
- Grootendorst, P.V., Feeny, D.H., y Furlong, W. (1997). Does it matter whom and how you ask? Inter- and intra-rater agreement in the Ontario Health Survey. *Journal of Clinical Epidemiology*, 50, 127-135.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., y Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.
- Groves, R.M. (1989). *Survey errors and survey costs*. Nueva York: John Wiley & Sons.
- Hambleton, R.K., y Pitoniak, M.J. (2002). Testing and measurement: Advances in item response theory and selected testing practices. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology* (3ª Edición- pp. 517-561). New York: John Wiley & Sons.
- Hidalgo, M. D., y Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, y B. McGaw (Eds.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier - Science & Technology.
- Jabine, T.B., Straf, M.L., Tanur, J.M., y Tourangeau, R. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. Washington, DC: National Academy Press.
- Johnson, B., y Christensen, L. (2008). *Educational research quantitative, qualitative, and mixed approaches*. Thousand Oaks, CA: Sage
- Kalaycioglu, D.B., y Berberoglu, G. (2011). Differential Item Functioning Analysis of the Science and Mathematics Items in the University Entrance

- Examinations in Turkey. *Journal of psychoeducational assessment*, 29 (5), 467-478.
- Kane, M.T. (1992). An argument-base approach to validity. *Psychological Bulletin*, 112 (3), 527-535.
- Krosnick, J.A. (1999). Survey Research. *Annual Review of Psychology*, 50, 537-567.
- Lewis, T.T., Yang, F.M., Jacobs, E.A., y Fitchett, G. (2012). Racial/Ethnic Differences in Responses to the Everyday Discrimination Scale: A Differential Item Functioning Analysis. *American Journal of Epidemiology*, 175 (5), 391-401.
- Magaziner, J., Bassett, S.S., Hebel, J.R., y Gruber-Maldini, A. (1996). Use of Proxies to Measure Health and Functional Status in Epidemiologic Studies of Community-dwelling Women Aged 65 Years and Older. *American Journal of Epidemiology*, 143, 283-292.
- Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics* 7, 105-118.
- Messick, S. (1989). Validity. En R.L. Linn (ed.): *Educational measurement*. (pp.13-103). New York: MacMillan.
- Miller, K. (2007). *Design and Analysis of Cognitive Interviews for Cross-National Testing*. Congreso de la European Survey Research Association. Praga, Checoslovaquia.
- Miller, K., Chepp, V., Willson, S., y Padilla, J. L. (2012, en prensa). Cognitive interviewing methodology: A sociological approach for survey question evaluation. NJ: John Wiley & Sons.
- Millsap, R. E., y Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 17, 297-334.

- Ministerio de Sanidad y Consumo (2006). Encuesta Nacional de Salud de España 2006. Madrid: Ministerio de Sanidad y Consumo.
- Ministerio de Sanidad y Consumo (2007). Encuesta de Discapacidad, Autonomía personal y situaciones de Dependencia de España 2006. Madrid: Ministerio de Sanidad y Consumo.
- Organisation for Economic Co-operation and Development. (2009b). *PISA 2006. Technical Report*. París: OECD publications.
- Oksenberg, L., Cannell, C., y Kalton, G. (1991). New Strategies for Pretesting Survey Questions. *Journal of Official Statistics*, 7, 349-365.
- Ongena, I.P., y Dijkstra, W. (2006). Methods of Behavior Coding of Survey Interviews. *Journal of Official Statistics*, 22, 419-451.
- Pickard, A. S., Johnson, J. A., Feeny, D. H., Ashfaq, M. D., Carriere, K. C., y Abdul, M. N. (2004). Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and health utilities index. *Stroke*, 35, 607-612.
- Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., y Singer, E. (2004). *Methods for Testing and Evaluating Survey Questionnaires*. Nueva York: John Wiley.
- Rajmil, L., Fernández, E., Gispert, R., Rúa, M., Glutting, J.P., Plasencia, A., y Segura, A. (1999). Influence of proxy respondents in childrens' health interview survey. *Journal of Epidemiology and Community Health*, 53, 38-42.
- Reichardt, C. S., y Rallis, S. F. (1994). Qualitative and quantitative inquiries are not incompatible: A call for a new partnership. *New Directions for Program Evaluation* 61, 85-91.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5-24.

- Sireci, S.G. (2009). Packing and Unpacking Sources of Validity Evidence: History Repeats Itself Again. In R. Lissitz (Ed.). *The Concept of Validity: Revisions, New Directions and Applications* (19-37). Charlotte, NC: Information Age Publishing Inc.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., y Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27, 53-75.
- Tafforeau, J., Lopez, M., Tolonen, A., Scheidt-Nave, C., y Tinto, A. (2006). *Guidelines for the development and criteria for the adoption of Health Survey instruments*. Eurostat Working Papers and Studies.
- Tashakkori, A., y Creswell, J.W. (2007). Exploring the nature of research questions in mixed methods research. *Journal of Mixed Methods Research*, 1, 207-211.
- Tashakkori, A., y Teddlie, C. (1998). *Mixed methodology: combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.
- Tashakkori, A., y Teddlie, C. (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. Thousand Oaks: Sage.
- Tourangeau, R. (1984). Cognitive science and survey methods: a cognitive perspective. En: T. Jabine, M. Straf, J. Tanur y R. Tourangeau (Eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between the Disciplines* (pp. 73-100). Washington, DC: National Academy Press.
- Todorov, A., y Kirchner, C. (2000). Bias in proxies. Reports of disability: Data from the National Health Interview Survey on Disability. *American Journal of Public Health*, 90, 1248-1253.
- Van de Vijver, F. J. R. (1998). *Towards a theory of bias and equivalence*. Mannheim: ZUMA.

- Van den Noortgate, W. V., De Boeck, P., y Meulders, M. (2003). Cross-Classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28, 369-386.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks: Sage Publications.
- Willis, G.B., DeMaio, T.J., y Harris-Kojetin, B. (1999). Is the bandwagon headed to the methodological promised land? Evaluating the validity of cognitive interviewing techniques. En M. Sirken, D. Herrmann, S. Schechter, N. Schwarz, J. Tanur y R. Tourangeau (Eds.), *Cognitive and survey research* (pp. 133-153). Nueva York: John Wiley & Sons.
- Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, 4(2), 223-233.
- Zumbo, B. D. (2009). Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice. In R.W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, (pp. 65-82). NC: IAP - Information Age Publishing.
- Zumbo, B.D., y Shear B.R. (2011). *The Concept of Validity & Some Novel Validation Methods*. Paper presented at the 42nd Annual Meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

ANEXOS



APPENDIX

Analysis of Quality of Proxy Questions in Health Surveys by Behavior Coding

Western Journal of Nursing Research
XX(X) 1–16
© The Author(s) 2010
Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>
DOI: 10.1177/0193945910388049
<http://wjn.sagepub.com>



Isabel Benítez¹, José Luis Padilla¹,
and Yfke Ongena²

Abstract

The aim of this study is to show how to analyze the quality of questions for proxy informants by means of behavior coding. Proxy questions can undermine survey data quality because of the fact that proxies respond to questions on behalf of other people. Behavior coding can improve questions by analyzing interviewer–respondent interactions. Twenty-nine proxies participated in the pretesting of a disability questionnaire. The questionnaire includes 11 questions related to daily-life limitations as a result of health problems. Interviewer–proxy interactions were coded and analyzed by means of Sequence Viewer program. The percentages, from a methodological perspective, of ideal “question-and-answer” sequences varied from 28% to 76% throughout the 11 questions analyzed. The results obtained pointed out the necessity of reviewing some of the proxy questions analyzed. Behavior coding can improve the quality of proxy questions in health surveys when proxy informants are surveyed.

Keywords

behavior coding, proxy informants, health surveys

¹University of Granada, Granada, Spain

²University of Twente, Enschede, Netherlands

Corresponding Author:

José Luis Padilla, Dpto. Psicología Social y Metodología de las CC. Comportamiento,
University of Granada, Campus de Cartuja, 18071 Granada, Spain
Email: jpadilla@ugr.es

The use of indirect informants or “proxies” to obtain information about other household members is common in household health surveys (Duncan et al., 2002; Magaziner, Speaar, Hebel, & Gruber-Baldini, 1996; Pickard et al., 2004; Schwarz & Wellens, 1997). A *proxy* is a person who answers survey questions about the health conditions of other people, whereas *self-reporters* answer about themselves. Proxy reporters are often used to fill in the designated household rosters in household surveys. The answers of the proxy reporter determine the eligibility of other household members to respond to other sections or questionnaires used in the survey. In addition, the use of proxies is frequent in medical processes or disease evaluations. Proxies have been used in an evaluation of the quality of life for reporting on patients with communication difficulties resulting from cerebral injuries (Sneeuw, Aaronson, de Haan, & Limburg, 1997). In turn, proxy and self-reporter answers have been compared to evaluate the validity of a questionnaire for patients who have suffered a stroke (Teixeira-Salmela, Devaraj, & Olney, 2007).

The guidelines and quality criterion for the design of health surveys prepared by the Eurostat Task Force, summing up the present consensus about the use of proxy reporters, indicates that the use of proxy-reporters should be limited only to cases in which (a) people are incapable of responding to questions, due to serious health problems (e.g., dementia, physical or severe mental disability), or (b) to those for whom it is not possible to interview for legal reasons (i.e., minors; Tourangeau, 2003). Nevertheless, using proxies is a common practice in national statistical institutes for the accomplishment of health surveys in numerous countries. The Health Examination Survey database (Koponen & Aromaa, 2001) promoted by the Scientific Institute of Public Health includes information provided by 34 countries from surveys in which proxy informants have been used, including Belgium (Health Interview Survey), Czech Republic (Labour Force Sample Survey), France (Survey on Household Living Conditions), Holland (Continuous Quality of Life Survey), and Spain (Survey of Disability, Personal Autonomy and Dependence Situations). Proxy informants were also used in the National Health Interview Survey on Disability for the National Center for Health Statistics in the United States (Todorov & Kirchner, 2000).

Few studies have looked at how to increase the quality of the answers provided by proxies, in spite of the fact that the use of proxies has been traditionally considered a threat to the quality of survey data (Ávila-Funes, Gray-Donald, & Payette, 2006). One of these studies evaluated the bias in the proxy answers by means of the National Health Interview Survey on Disability carried out in New York. The results showed that proxies used different response strategies than self-reporters (Todorov & Kirchner, 2000). Another study, in which

proxy answers were evaluated on a scale of cerebral injury impact, showed that proxy and patient evaluations are more consistent when they evaluate observable and specific behaviors, whereas the agreement decreased when the proxy informants made subjective judgments (Duncan et al., 2002). On the other hand, while evaluating the quality of life in patients who have suffered cerebral injuries, it was found that the proxies' evaluations were sensitive to the differences in the patients' functionality (Sneeuw et al., 1997).

Evaluating proxy responses is especially challenging in disability survey contexts because, according to the World Health Organization (WHO, 2009) definition of disability, the classification of a person as having or not having a disability is a subjective judgment as it depends on the interaction between social conventions, individuals, cultural norms, expectations, etc. Therefore, the responses to the questions about whether a person has a disability could vary according to the type of informant (self-reporter vs. proxy), as a result of potential differences between both norms and expectations, but not necessarily as a result of objective information.

Pretest methods can be helpful in improving survey questions. The general objective of pretest methods is the identification of the causes of errors in surveys by means of the analysis of the events occurring during the "question and answer" process (Willis, 2005). Behavior coding is one of the pretest methods used by survey methodologists, either on its own or in combination with other pretest methods such as cognitive interviewing, focus groups, or speech analyses, to optimize the question drafting and the questionnaire design (Presser et al., 2004). In contrast with such pretest methods, behavior coding provides systematic, objective, and replicable results (Groves et al., 2004).

The behavior coding method was developed in the 1960s by Charles Canell to evaluate both the questions and the interviewer behavior (Canell, Fowler, & Marquis, 1968). Behavior coding is based on the rationale that the interviewer's and respondent's behaviors provide information about potential problems with survey questions related to question phrasing and to questionnaire design by systematically observing the interviewer–respondent interaction (Blair & Srinath, 2008). Moreover, behavior coding allows survey researchers to evaluate the quality of survey questions aimed at specific respondent groups defined by characteristics such as age, educational level, or gender. Nevertheless, little attention has been given to questions designed for respondents with different roles (self-reporter or proxy) in the interview process.

The aim of this study is to show how to analyze the quality of proxy questions by means of behavior coding in a health survey. In this study, the adequacy of the questions to be answered for proxies will be also discussed.

Method

Participants

Twenty-nine proxy informants, 13 men and 16 women, with an age average of 31.06 years took part in the pretest of a disability questionnaire. The educational level of participants was balanced (14 participants with less than 14 years of schooling and 15 participants with more than 14 years of schooling). The sample size of the study is within the interval (15-50) recommended by several authors to maximize the usefulness of results provided by the behavior coding method (Blair & Srinath, 2008).

All participants were Spanish and they provided information only about people with whom they lived and had a direct familiar relationship, for instance, parents, partners, brothers, or sisters. The selection was carried out with regard to various requirements that determine if the participant was eligible, that is to say, they had the same characteristics of the target population of the future health survey in which the tested questions in this study would be administrated.

It was also confirmed that the participants had not previously taken part in a survey pretest. The participants were contacted via associations for disabled person support, and they received 30 euros for taking part in the study.

Materials

The people responsible for carrying out the interviews used interview protocols during the pretest that included demographic questions and 11 “target” questions. The target questions were the ones selected to be analyzed during the pretest by means of behavior coding. These were selected by experts, who evaluated the questionnaire and identified questions that could present difficulties. These experts had a long experience in the field of health surveys and survey methodology. Table 1 shows the 11 questions to be analyzed by means of the behavior coding method.

Procedure

The interviews, in which the questionnaire with the target questions was applied, were conducted by two trained and experienced interviewers (one male and one female). They were specifically instructed to ask target questions as the questions were worded in the questionnaire. The interviews were conducted in a laboratory specially equipped to perform cognitive pretesting. Confidentiality

Table 1. Selected Questions From the Disability Questionnaire

Target Questions

-
- Q.1. Is there any person in your home who has been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year.
- Q.2. Is there any person in your home who has serious difficulty speaking in an understandable manner and saying meaningful phrases without help?
- Q.3. Is there any person in your home who has serious difficulty understanding the meaning of what others say without help?
- Q.4. Is there any person in your home who has serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille.
- Q.5. As a result of problems of a cognitive or intellectual nature, is there any person in your home who has serious difficulty when intentionally using the senses? For example, paying visual attention, listening attentively, etc.
- Q.6. As a result of problems of a cognitive or intellectual nature, is there any person in your home who has serious difficulty learning to read, write, count (or calculate), copy or difficulty learning to use everyday utensils?
- Q.7. Is there any person in your home who has serious difficulty moving one's body from one place to another without changing position, without help and without supervision? For example, going from sitting on the bed to sitting on a chair.
- Q.8. Is there any person in your home who has serious difficulty changing posture without help and without supervision? For example, getting up from a chair, lying down on the bed, kneeling down, etc. Exclude the action of moving one's body posed in the previous question.
- Q.9. Is there any person in your home who has serious difficulty showing other people affection, respect or transmitting feelings, including physical contact such as kisses, caresses, etc.?
- Q.10. Is there any person in your home who has serious difficulty forming and maintaining family relationships?
- Q.11. Is there any person in your home who has serious difficulty forming and maintaining sentimental or sexual relationships with a partner?
-

and the exclusive use of the information for research purposes were ensured. Having obtained the respondents' consent, the interviews were audio and video recorded. The interviews were transcribed and two coders used the transcripts and recordings to systematically classify the interviewer and respondent behaviors. The two coders worked independently, and once first classifications were made, they met to analyze discrepancies and reach an agreement.

Table 2. Categories for the Classification of Respondents' Behaviors

Codes	Meaning
During question reading	
Request clarification	Explicit expression for indicating problems in the comprehension of the concepts included in the question or in the task comprehension.
Interruption	The respondent stops the question reading (to request clarification or to answer).
Answer	
Mismatch answer	The response is adequate but is not exactly worded as any of the answer options.
Invalid answer	The response is not related to the question.
Don't know answer	The respondent does not know how to respond.
Qualified answer	The response indicates uncertainty.
Adequate answer	The response fits the objective of the question.

Verbal Behavior Coding

The behavior coding was done by means of the Sequence Viewer program (Dijkstra, 2008). Coders were also trained by experts in Sequence Viewer program. This program provides information about possible problems with the content or the format of the questionnaire, by systematic classification of behaviors occurring during the interview. The analysis begins with the division of the transcripts into sequences. A sequence starts with the reading of a question and ends when the reading of the following question starts (Dijkstra, 1999). The sequences are analyzed by assigning different codes depending on the behaviors occurring during the interviewer–respondent interaction. For example, while interviewers are asking questions, respondents can ask for explanations or extra information (coded as “request for clarification”), and respondents can interrupt the interviewer giving their answers to the question before the interviewer has finished reading or making comments (coded as “interruption”). Answers given by the respondent after interviewers have finished reading the question can be classified in different ways, of which the classification realized by Oksenberg, Cannell, and Kalton (1991) is the most commonly used. This classification has been extended by authors like Van der Zouwen and Smit (2004), Forsyth, Levin, and Fisher (1999), and Ongena (2005). Table 2 shows the coding scheme used in this study, which is primarily based on the classification by Oksenberg et al. (1991).

To evaluate the quality of proxy questions, codes were used in the study as indicators of response accuracy. A scale of accuracy was developed, using extremes represented by the codes “adequate answers” (being the most accurate) and “invalid answers” (being the most inaccurate). The intermediate categories were defined as “mismatch answer,” “qualified answer,” and “don’t know answer.”

Depending on the combination of codes assigned, sequences are classified as “paradigmatic sequences,” “nonparadigmatic–nonproblematic sequences,” and “nonparadigmatic–problematic sequences.” A *paradigmatic sequence* is defined as the ideal sequence during the question-and-answer process. An ideal sequence is that in which the delivery of the question is identical to that indicated in the interview protocol, the respondent’s answer is adequate, and the interviewer recognizes the answer as being adequate (Ongena & Dijkstra, 2006). A *nonparadigmatic sequence* is problematic or nonproblematic depending on whether the type of behavior occurring is considered to be a problematic influence on the data. In this study, the occurrence of mismatch answers, invalid answers, don’t know answers, qualified answers, and requests for clarification all classify the sequence as a problematic sequence. A sequence is classified as nonparadigmatic–nonproblematic when deviations occur that are not problematic (e.g., interruptions).

Sequences are classified considering the codes assigned to each behavior that occurred during the sequence. For example, the occurrence of the behavior “request clarification” causes a sequence to become nonparadigmatic although the respondent’s answers were adequate.

Once the sequences were classified, a frequency analysis was performed that consisted first of calculating the frequencies of each type of sequence followed by calculating the rate of the occurrence of problematic answers. When 15% or more of a question’s administrations show one or more problematic interactions, it is a widely accepted criterion for determining the question to be flawed (Blair & Srinath, 2008). On the other hand, if the percentage of nonparadigmatic sequences is considered, questions in which the percentage is greater than 60% must be checked (Van der Zouwen & Dijkstra, 2002). Analysis of 319 sequences (i.e., 11 questions \times 29 respondents) was conducted using both criteria to illustrate the use of behavior coding in the study.

Results

For the present analysis, 319 sequences (i.e., 11 questions \times 29 respondents) were taken into account.

Table 3. Frequencies and Percentages of Each Type of Sequence

Target Questions	Type of Sequence					
	Paradigmatic Sequence		Nonparadigmatic–Nonproblematic Sequence		Nonparadigmatic–Problematic Sequence	
	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>
Q.1. Habitual activities	28	8	21	6	52	15
Q.2. Speak	76	22	3	1	21	6
Q.3. Understand	66	19	24	7	10	3
Q.4. Use the phone	41	12	17	5	41	12
Q.5. Use the senses	66	19	10	3	24	7
Q.6. Learn	69	20	17	5	14	4
Q.7. Move the body	76	22	10	3	14	4
Q.8. Change posture	66	19	14	4	21	6
Q.9. Show affection	76	22	14	4	10	3
Q.10. Family relationships	72	21	10	3	17	5
Q.11. Sentimental relationships	55	16	34	10	10	3

Types of Sequence

First, the behavior coding analyses showed the frequency of occurrence of each type of sequence produced by the proxy informants. Table 3 shows the percentage of occurrence of each type of sequence for each target question.

As Table 3 shows, the percentage of paradigmatic sequences, that is to say, ideal sequences from the methodological point of view, ranges between 28% and 76% for the target questions. The Cramer's *V* statistic (.2762) indicates a low association between the type of sequence and the target question analyzed. Target Question 1 showed the highest percentage of nonparadigmatic–problematic sequences (52%). Following the usual criteria, Question 1 was recommended for checking, because 72% of the sequences were classified as nonparadigmatic. This high percentage could be due to the content of

Table 4. Frequencies and Percentages of Answer Category Codes

Target Questions	Codes									
	Mismatch Answer		Invalid Answer		Don't Know Answer		Qualified Answer		Adequate Answer	
	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>
Q.1. Habitual activities	45	13	0	0	0	0	6	2	90	26
Q.2. Speak	0	0	6	2	0	0	3	1	90	26
Q.3. Understand	0	0	7	2	0	0	0	0	97	28
Q.4. Use phone	14	4	21	6	3	1	7	2	76	22
Q.5. Use senses	7	2	7	2	3	1	7	2	93	27
Q.6. Learn	10	3	3	1	0	0	3	1	100	29
Q.7. Move the body	10	3	3	1	0	0	0	0	100	29
Q.8. Change posture	10	3	10	3	3	1	0	0	90	26
Q.9. Show affection	3	1	7	2	0	0	3	1	93	27
Q.10. Family relationships	3	1	3	1	0	0	0	0	97	28
Q.11. Sentimental relationships	3	1	3	1	3	1	10	3	90	26
Cramer's <i>V</i>	.403		.209		.149		.180		.239	

the question, which is more general and ambiguous than the rest of the target questions.

Codes for Proxy Responses

In this study, we were particularly interested in deviations produced by proxies. Table 4 shows the percentages of adequate and (four types of) inadequate answers for the 11 questions of the disability questionnaire. The percentages per row add up to more than 100% because multiple behaviors can occur in one sequence. For example, the respondent can change the answer once it is coded as “invalid answer.”

As Table 4 shows, the Cramer's *V* values reveal a low association between the type of answer produced by the participants and the question analyzed in all the cases except for the code mismatch answer. This code

shows the highest percentage of occurrence for the set of target questions. Question 1 achieved the highest percentage of mismatch answers (45%). The following example represents a situation in which a mismatch answer was produced:

Interviewer: Is there any person in your home who has been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year.

Yes, seriously limited; yes, limited but no seriously; not.

Respondent: Yes

The answer given by the respondent was coded as a “mismatch answer” because it does not fit to any of the response alternatives offered. The high percentage of mismatch answers found in question 1 might be due to respondents’ understanding it as a yes/no question without considering the three response alternatives offered.

Question 4 achieved the highest percentage of invalid answers (21%) and the lowest percentage of adequate answers (76%). The following example represents an invalid answer found in Question 4:

Interviewer: Is there any person in your home who has serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille.

Respondent: In my home, nobody knows how to use the machines for writing in Braille.

The answer was coded as an “invalid answer” because its content is not related to the intended objective of the question.

Finally, in Question 11, 10% of answers were registered as qualified. An example from the interviews illustrates the meaning of the code qualified answer.

Interviewer: Is there any person in your home who has serious difficulty forming and maintaining sentimental or sexual relationships with a partner?

Respondent: I don’t think so.

In Question 11, the high percentage of qualified answers may indicate that the proxies have doubts when responding to questions on personal topics

Table 5. Frequencies and Percentages of Difficulty Indicator Codes During the Question Reading

Target Questions	Codes			
	Request Clarification		Interruption	
	%	<i>n</i>	%	<i>n</i>
Q.1. Habitual activities	7	2	3	1
Q.2. Speak	3	1	3	1
Q.3. Understand	0	0	3	1
Q.4. Use the phone	14	4	0	0
Q.5. Use the senses	3	1	0	0
Q.6. Learn	3	1	0	0
Q.7. Move the body	0	0	0	0
Q.8. Change posture	7	2	0	0
Q.9. Show affection	3	1	0	0
Q.10. Family relationships	7	2	0	0
Q.11. Sentimental relationships	3	1	0	0
Cramer's <i>V</i>	.1740		.1591	

such as sexual or personal relationships. Nevertheless, this leaves the interviewer with a dilemma: Should she further probe for an unqualified answer, or just accept the answer as given? In some cases, this is not necessary, as respondents may spontaneously repair their qualified answer by giving an unqualified adequate answer afterwards.

Difficulty Indicators When Asking Questions

“Request clarification” and “interruption” are codes commonly used in behavior coding as indicators to identify difficulties while interviewers are asking questions. Table 5 shows the frequencies of both codes in the 11 target questions.

As Table 5 shows, the Cramer's *V* values reflect a low association between the behaviors produced by the participants and the target question analyzed. Requests for clarification occurred most frequently with Question 4. Questions 8 and 10 also showed a high percentage in the appearance of this code. The following example demonstrates the occurrence of such a request clarification:

Interviewer: Is there any person in your home who has serious difficulty forming and maintaining family relationships?

Respondent: Family relationships?

Interruptions were coded to some extent in Question 2. The excerpt illustrates an interruption found in Question 2:

Interviewer: Is there any person in your home who has serious difficulty speaking . . .

Respondent: Yes

Interviewer: . . . in an understandable manner and saying meaningful phrases without help?

In this specific case, difficulties could arise from an interruption, since the respondent is answering the question before hearing all the elements that have to be considered (Van der Zouwen & Dijkstra, 2002).

Discussion

The aim of the study was to illustrate how to analyze the quality of the questions intended for proxy respondents in a health survey by means of behavior coding. The results from the behavior coding application to the disability questionnaire pretested in the study allowed the quality of the proxy questions to be analyzed.

The general results showed percentages of paradigmatic sequences between 28% and 76% for the set of 11 target questions. Only Question 1, habitual activities, achieved more than 60% of nonparadigmatic sequences.

The results highlighted some questions to be checked or in which it was necessary to examine the proxies' behavior in detail. These problems might be due to the characteristics of the questions, or to the role represented by the informants. For example, Question 1 (habitual activities) is worded as a yes/no question while three alternatives are offered to the respondent. This is a problem that is common in survey questionnaire design (Ongena, 2003). In addition, two of the three options are positive ("Yes, seriously limited" and "Yes, limited but no seriously"), and one is negative ("Not"). Thus, researchers find an adequate answer in cases in which the respondents' answer is negative, but a large percentage of mismatch answers when the respondents' answers are positive but they replied with a simple yes. Assessing how serious the limitation was and distinguishing between the affirmative alternatives can be a difficult task for proxies. On the other hand, proxy behavior

could cause measurement error because either proxies focus on aspects that are not the aim of the question (Question 4 on the use of the telephone) or they face nonobservable or sensitive topics (Question 10 on family relationships). Possible impact of demographics, such as educational level, degree of family relationship, and so on, on proxy questions was not specifically addressed in our study because of its particular design, which can be considered a limitation.

When using proxies, survey researchers consider several factors. The difficulty of the task and the motivation for responding to questions could be different for self-reporters than for proxies. In addition, proxy respondents may have less information available in their episodic memory (Schwarz & Wellens, 1997). More studies focused on comparing the proxies and self-reporter behavior are necessary, as well as evaluating the convergence between the answers provided by both types of informants. Future research may address these topics.

Respondent behavior can be studied from multiple perspectives, including more qualitatively oriented studies. For example, Collins, Shattell, and Thomas (2005) address how to deal with potentially problematic interviewee behaviors, such as flattery, filtration, or statements indicative of social desirability response bias for qualitative research. Behavior coding as a method provides a systematic approach to analyze interviewer and respondent behavior, is flexible, and offers the possibility of obtaining qualitative and quantitative information that help survey methodologists improve survey data quality. In comparison with other pretest methods, behavior coding is focused on the participant's behavior. The assumption behind behavior coding is that the interviewer–respondent interaction can provide very useful information about potential problems with question phrasing and questionnaire design. This information allows survey researchers to identify questions with a high percentage of “problematic behaviors” as questions that should be revised.

Behavior coding also presents some limitations. For example, it is possible that a respondent gives an adequate answer although he has not understood the real sense of the question. In fact, there may be a gap between respondents' observed behaviors and their understanding of the key concepts in the questions. Combining behavior coding and cognitive interviewing can resolve that gap. Future research in the pretest methods field should address how to combine evidence provided by different pretest methods.

On the other hand, it is necessary to reach a greater consensus about the criteria used to check the questions on the basis of the results obtained by means of the behavior coding. A review of studies in which behavior coding was used found that some authors consider those questions problematic in which the percentage of adequate answers was lower than 85% and some

other authors when that was lower than 90%, whereas others focused on the percentage of inadequate answers, recommending to review the questions in which the percentage is greater than 15% (Van der Zouwen & Smit, 2004). The criteria used can cause changes in the conclusions drawn because, for example, an adequate answer can occur after an inadequate answer. If an inadequate answer criterion is used, a question can be eliminated although a high percentage of final adequate answers has been reached.

Behavior coding has shown its usefulness in evaluating the quality of questions designed for proxy informants by providing detailed information about the participants' behavior and facilitating the detection of possible sources of measurement error. However, more research is needed to find out the causes of question problems identified by coding behavior and their consequences when results of behavior coding studies are applied in survey questionnaire design, especially when proxy questions are included in the survey questionnaire. Nevertheless, as Oksenberg et al. (1991) highlight, there is convincing evidence of the usefulness of behavior coding in improving the quality of survey questions providing quantitative, systematic, and replicable results.

Declaration of Conflicting Interests

The authors declared no conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received financial support for the research and/or authorship of this article: Funded by Spain's Ministry of Science and Innovation under the European Regional Development Fund (project no. PSI2009-07280), and the Programa de Incentivos a Proyectos de Investigación de Excelencia, Consejería de Innovación y Ciencia, Junta de Andalucía (SEJ-5188).

References

- Ávila-Funes, J. A., Gray-Donald, K., & Payette, H. (2006). Medición de las capacidades físicas de adultos mayores de Quebec: un análisis secundario del estudio NuAge [Measurement of physical capacities in the elderly: A secondary analysis of the Quebec longitudinal study NuAge]. *Salud Pública en México*, *48*, 446-454.
- Blair, J., & Srinath, K. P. (2008). A note on sample size for behavior coding pretests. *Field Methods*, *20*, 85-95.
- Cannell, C. F., Fowler, F. J., & Marquis, K. H. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting of household interviews. *Vital and Health Statistics*, *2*(26), 1-65.

- Collins, M., Shattell, M., & Thomas, S. P. (2005). Problematic interviewee behaviors in qualitative research. *Western Journal of Nursing Research, 27*, 188-199.
- Dijkstra, W. (1999). A new method for studying verbal interactions in survey interviews. *Journal of Official Statistics, 15*, 67-85.
- Dijkstra, W. (2008). Sequence viewer (version 4.4a) [Computer software]. Free University of Amsterdam, Netherlands.
- Duncan, P., Min Lai, S., Tyler, D., Perera, S., Reker, D. M., & Studenski, S. (2002). Evaluation of proxy responses to the stroke impact scale. *Stroke, 33*, 2593-2599.
- Forsyth, B., Levin, K., & Fisher, S. (1999). Test of an appraisal method for establishment survey questionnaires. In *Proceedings of the ASA Section on Survey Research Methods* (pp. 145-149). Alexandria, VA: American Statistical Association.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.
- Koponen, P., & Aromaa, A. (2001). Health examination surveys (HES) in the European Union: Review of literature and inventory of surveys in the EU/EFTA Member States. Retrieved from the Scientific Institute of Public Health website <http://www.iph.fgov.be/keywords.asp?Lang=EN&ReportID=2107>
- Magaziner, J., Speaar, S., Hebel, J. R., & Gruber-Baldini, A. (1996). Use of "proxies" to measure health and functional status in epidemiologic studies of community-dwelling women. *American Journal of Epidemiology, 143*, 283-292.
- Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics, 7*, 349-365.
- Ongena, Y. P. (2003, June). *Pre-testing the ESS-questionnaire using interaction analysis*. Paper presented at the European Social Survey CCT meeting, Sociaal en Cultureel Planbureau, The Hague, Netherlands.
- Ongena, Y. P. (2005). *Interviewer and respondent interaction in survey interviews* (Unpublished doctoral dissertation). Vrije Universiteit, Amsterdam, Netherlands.
- Ongena, Y. P., & Dijkstra, W. (2006). Methods of behavior coding of survey interviews. *Journal of Official Statistics, 22*, 419-451.
- Pickard, A. S., Johnson, J. A., Feeny, D. H., Ashfaq, M. D., Carriere, K. C., & Abdul, M. N. (2004). Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and health utilities index. *Stroke, 35*, 607-612.
- Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires*. New York, NY: Wiley-Interscience.
- Schwarz, N., & Wellens, T. (1997). Cognitive dynamics of proxy responding: The diverging perspectives of actors and observers. *Journal of Official Statistics, 13*, 159-174.

- Sneeuw, K. C., Aaronson, N. K., de Haan, R. J., & Limburg, M. (1997). Assessing quality of life after stroke: The value and limitations of proxy ratings. *Stroke, 28*, 1541-1549.
- Teixeira-Salmela, L. F., Devaraj, R., & Olney, S. J. (2007). Validation of the human activity profile in stroke: A comparison of observed, proxy and self-reported scores. *Disability Rehabilitation, 29*, 1518-1524.
- Todorov, A., & Kirchner, C. (2000). Bias in "proxies." Reports of disability: Data from the National Health Interview Survey on Disability. *American Journal of Public Health, 90*, 1248-1253.
- Tourangeau, R. (2003). Cognitive aspects of survey measurement and mismeasurement. *International Journal of Public Opinion Research, 15*, 3-7.
- Van der Zouwen, J., & Dijkstra, W. (2002). Testing questionnaires using interaction coding. In D. Maynard, H. Houtkoop-Steenstra, N. Schaeffer, & J. Van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the Survey interview* (pp. 427-447). New York, NY: John Wiley.
- Van der Zouwen, J., & Smit, J. H. (2004). Evaluating survey questions by analyzing patterns of behaviour codes and transcripts of question-answer sequences: A diagnostic approach. In S. Presser, J. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 109-130). New York, NY: John Wiley.
- Willis, G. B. (2005). *Cognitive interviewing*. Thousand Oaks, CA: Sage.
- World Health Organization. (2009). *ICIDH-2: International Classification of Functioning, Disability and Health (ICF)*. Geneva, Switzerland: Author.

Evaluation of the convergence between “self-reporters” and “proxies” in a disability questionnaire by means of behaviour coding method

Isabel Benítez Baena · Jose Luis Padilla García · Yfke Ongena

Published online: 25 March 2011
© Springer Science+Business Media B.V. 2011

Abstract Household surveys often require including proxy reporters to obtain information about other household members who cannot be interviewed. The participation of proxies can undermine survey data quality due to the fact that proxies must respond to questions thinking about other people. The objectives of the present study were to analyze the behaviour of proxy reporters and evaluate the convergence between the answers given by proxies and self-reporters by means of behaviour coding. This improves the evaluation of convergence, since only adequate (i.e., interpretable) answers given by both types of informant are taken into account. Responses to a disability questionnaire employed by an official statistical institute were analyzed. The questionnaire includes 11 questions about different limitations related to everyday activities. 16 self-reporter and 16 proxies formed 16 couples whose members lived together and supported a direct family relation. The results show a high percentage (52%) of convergence between both types of informant, although fluctuating across the questions and the couples. Proxies showed relatively more adequate behaviour during the interaction than self-reporters. From this we conclude that proxies can be considered at least as good informants as self-reporters from an interviewer-respondent interaction perspective. Future research should address the impact of proxy responses on survey validity.

Keywords Proxies · Pretest · Behaviour coding · Convergence evaluation · Disability questionnaire

I. Benítez Baena · J. L. Padilla García (✉)
Department of Social Psychology and Methodology of Behavioral Sciences,
University of Granada, Campus de Cartuja, 18071 Granada, Spain
e-mail: jpadilla@ugr.es

Y. Ongena
Department of Communication Studies,
University of Groningen, Groningen, The Netherlands

1 Introduction

One of the most notable characteristics of the design of many household surveys is the use of proxy reporters to obtain information about other household members. A “proxy” is a person who answers the survey thinking about another person of their household or environment, whereas a “self-reporter” answers thinking about himself. Proxy reporters are often employed to fill in the designated household rosters in household surveys. The answers of the proxy reporter determine the eligibility of the other household members for responding to other sections or questionnaires used in the survey. The decision to use proxy reporters is the result of weighing up costs, sampling errors and response errors (Sunghee et al. 2007). Nonetheless, the reduction of costs can be outweighed by the increase in measurement errors when compared with self-response reporting.

The use of proxies is a common practice in household surveys carried out by official statistical institutes, despite the fact that the document on guidelines and quality criterion for the design of health surveys prepared by the *Eurostat Task Force* for the design of health surveys, summarizing the consensus among survey researchers, discourages interviewing proxies. The consensus among survey researchers indicates that the use of “proxies” should be limited only to ‘replacing’ people who are incapable of responding to questions, due to serious health problems (e.g., dementia, physical or severe mental disability, etc.) or those for whom it is not possible to interview for legal reasons, for example minors (Tafforeau et al. 2006).

Many studies have investigated the influence on survey data quality of the proxies’ characteristics, such as age, gender, educational level, level of income, and the relationship with the self-reporter. For example, Magaziner et al. (1996) found a high degree of agreement between proxies and self-reporters who live together. That study, which included differences in the type of information requested, shows that proxies were able to accurately report on health and observable functioning, such as physical or daily tasks, chronic conditions, etc. Nevertheless, the information provided by the proxies about the symptoms of health (frequently not observable and not discussed with others) was less precise. Some investigations also found a decrease in the precision when proxies report about psychosocial characteristics or symptoms (Pickard et al. 2004) or subjective areas like memory and thought, communication, emotion or behaviour (Duncan et al. 2002).

Not much is known about the effects of using proxies on data quality, but it is known that the answers sometimes differ from the answers provided by self-reporters. Pickard et al. (2004) evaluated the agreement between proxies and self-reporters by means of a questionnaire which evaluates the “quality-of-life” construct, finding systematic differences between the information given by both types of informant. In relation to the accuracy of the answers provided by each type of informant, some studies show that self-reporter’ answers are more precise than proxies’ answers (Loftus et al. 1992).

Schwarz and Wellens (1997) showed by means of several experiments that proxy reports show higher consistency than self-reporters. However, consistency does not necessarily mean more accuracy, as the information source for proxies may be biased. Schwarz and Wellens argue that proxies derive information to judge an answer from dispositional information (i.e., the personality and likes and dislikes of the person they are reporting on), whereas self-reporters are more likely to base their judgment on situational factors. Hence, questions on distant events and concerning lengthy reference periods will increase convergence of proxy and self-reporters, since for such questions self-reporters have less possibility of accessing episodic information on situational influences, and consequently, like the proxies, will use dispositional information.

Nevertheless, the use of proxies is necessary in surveys in which the self-reporter “cannot” be interviewed. This is often the case with surveys about health and well-being, or when respondents have different health conditions associated with age, for example a study in which proxies were used to retrieve the functional state of patients over 65 years old (Magaziner et al. 1996).

Evaluating proxy responses is especially challenging in disability survey contexts because, according to the World Health Organization (2009) definition of disability, the classification of a person as having or not having a disability is a subjective judgment as it depends on the interaction between social conventions, individuals, cultural norms, expectations, etc, that is to say, the level of disability is determined by the environment and its demands and not only by the diagnosed difficulties the person has. Therefore, the responses to the questions about whether or not a person has a disability could vary according to the type of informant (self-reporter vs. proxy), as a result of differences between both in norms or expectations, but not necessarily as a result of objective information.

Few studies have been focused on the potential sources of errors associated with the “role” assigned to the respondent. Todorov and Kirchner (2000) found a systematic evaluation bias in the proxies’ responses to the National Health Interview Survey on Disability. In this survey proxies were used in the cases when not all household members were available, in order to avoid having to return to the same households on repeated occasions.

Pretest methods, whose usefulness to optimize the information obtained by surveys have been widely proven, can be used to evaluate the influence of proxies on survey data quality. The general objective of pretest methods is the identification of the causes of errors in surveys by means of the analysis of the events happening during the “questions-and-answer” process (Tourangeau 2003). Tourangeau et al. (2000) formulated the most disseminated version of the “questions-and-answer” model with four sequential main phases: comprehension, retrieval, judgement and response selection. The extent to which respondents “pass” through each of these phases could be determined by the role assigned to the respondent: proxy vs. self-reporter. Thus, pretest methods could contribute to detecting differences in the cognitive process completed by proxies and self-reporters.

Among pretest methods, the behaviour coding method has proven its utility for providing information about the problems which can exist in relation to the formulation of the questions and the questionnaire format by means of the systematic observation of the interviewer-respondent interaction (DeMaio et al. 1998). Behaviour coding can detect problematic behaviours by classifying the events occurring during the interaction. In a general sense, behaviour coding allows the researcher to establish relations between the problematic behaviours identified and the respondent, interviewer and questionnaire characteristics.

Given the potential effects of the use of proxy reporters on measurement error in surveys, it is necessary to evaluate the convergence between proxies and self-reporters. A high convergence between both types of informant would lead to a higher confidence in using proxies when it is not possible to access self-reporting responses. Applying behaviour coding methods can improve the evaluation of convergence by analyzing only “adequate” answers, i.e., answers that are directly interpretable as an answer, given by both types of informant. The aim of the present study is to analyze the behaviour of proxy reporters and evaluate the convergence between proxies and self-reporters in a disability questionnaire by means of behaviour coding.

Table 1 Description of characteristics of the cognitive pretest participants

Subgroup	Gender		Age (Average)			
	Male	Female	16–25	26–45	46–60	+61
Self-reporters	6	10	1 (19)	4 (40.3)	8 (50.6)	3 (66)
Proxy reporters	7	9	6 (20.5)	6 (34.7)	4 (52)	0 (0)

2 Method

2.1 Participants

Sixteen couples, that is to say, 32 people (13 men and 19 women) participated in the cognitive pretest of the disabilities questionnaire included in a survey. The members of each couple were living together and they had a direct family relation. The selection was carried out with regard to various requirements that determine if the participant is “eligible” for a future administration of the survey. In all cases, the participants should have mastered functional Spanish, i.e., sufficient to manage everyday situations. With respect to demographic variables, the participants’ ages were between 16 and 80 years old. Lastly, it was checked that the participants had not previously taken part in a survey pretest. Table 1 presents the distribution of the demographic variables used for selecting the participants for the cognitive pretest.

2.2 Materials

Two versions of a disability questionnaire that differed in question wording depending on the respondent type, were used. The self-reporter version questions were addressed to the self-reporter with ‘you’, whereas in the proxy version questions, this ‘you’ was replaced by ‘any person in your home who’. Table 2 shows the 11 questions selected to be analyzed in the questionnaire pretest, called “target questions”, in the self-reporters version.

The interviewers used an interview protocol for performing the interviews. The interview protocol included the target questions together with the usual demographic questions. Interviews were recorded in video and audio, having previously obtained the respondents consent.

2.3 Procedure

During the recruitment phase, 16 people who met the necessary requirements to act as self-reporters were selected. These people were selected to be self-reporters regardless of whether or not they had any limitations when performing everyday activities. In addition, these people were requested to come to the interviews along with another household member, who would act as a proxy. The participants did not know in advance what their roles would be. Interviews were conducted individually and took place in cognitive laboratories equipped with video and audio recorders. Later behaviour coding was carried out using the transcripts and recordings of the interviews.

Table 2 Self-reporter target question

1.	Have you been limited in the performance of habitual activities due to a health problem? The limitation should have lasted or be expected to last more than 1 year.
2.	Do you have serious difficulty speaking in an understandable manner and pronouncing meaningful phrases without help?
3.	Do you have serious difficulty understanding the meaning of what others say without help?
4.	Do you have serious difficulty using the telephone or other devices or means of communication without help and without supervision? Include lip-reading and machines for writing in Braille.
5.	As a result of problems of a cognitive or intellectual nature, do you have serious difficulty when intentionally using the senses? For example, paying visual attention, listening attentively, etc.
6.	As a result of problems of a cognitive or intellectual nature, do you have serious difficulty learning to read, write, count (or calculate), copy or difficulty learning to use everyday utensils?
7.	Do you have serious difficulty moving your body from one place to another without changing position, without help and without supervision? For example, going from sitting on the bed to sitting on a chair.
8.	Do you have serious difficulty changing posture without help and without supervision? For example, getting up from a chair, lying down on the bed, kneeling down, etc. Exclude the action of moving one's body posed in the previous question.
9.	Do you have serious difficulty showing other people affection, respect or transmitting feelings including physical contact such as kisses, caresses, etc.?
10.	Do you have serious difficulty forming and maintaining family relationships?
11.	Do you have serious difficulty forming and maintaining sentimental or sexual relationships with a partner?

2.4 Analysis

The analysis was done using the program Sequence Viewer version 4.4.a (Dijkstra 2008). This program provides information about possible problems with the content or the format of the questionnaire, by classifying the behaviours occurring during the interview. The classification of the behaviour can be carried out depending on when it occurs: while the interviewers were asking the questions, or while the respondents were answering the questions.

While interviewers are asking questions, respondents can ask for explanations or extra information (coded as “request for clarification”), and respondents can interrupt the interviewer giving their answers to the question before the interviewer has finished reading or making comments (coded as “interruption”). Answers given by the respondent after interviewers finish reading the question can be classified in different ways, of which the classification realized by Oksenberg et al. (1991) is the most commonly used. This classification has been extended by authors like Van der Zouwen and Smit (2004); Forsyth et al. (1999), and Ongena (2005). Table 3 shows the version of the classification by Oksenberg, Cannell and Kalton used in this study.

Depending on the behaviours occurring, the sequence can be classified as paradigmatic, non paradigmatic-non problematic or non paradigmatic-problematic. A “paradigmatic sequence” is defined as the ideal sequence during the question-and-answer process (Schaeffer and Maynard 1996). In agreement with Ongena and Dijkstra (2006) an ideal sequence is that in which the delivery of the question is identical to that indicated in the script, the respondent's answer is adequate and the interviewer recognizes the answer as being adequate. A “non paradigmatic sequence” is problematic or non problematic depending on whether the type of behaviour occurring is considered to be a problematic influence on the data. In this study, the occurrence of mismatch answers, invalid answers, don't know answers, qualified answers and requests for clarification all classify the sequence as a problematic sequence.

Table 3 Responses categories to classify respondents' answers

Codes	Meaning
Problematic answers	
Mismatch answer	The response is adequate but doesn't coincide with any of the answer options
Invalid answer	The response is not related to the question
Don't know answer	The respondent did not know how to respond
Qualified answer	The response indicates uncertainty
Non problematic answer	
Adequate answer	The response fits the objective of the question

A sequence is classified as non paradigmatic- non problematic when deviations occur that are not problematic (for example, interruptions).

After this classification, a frequency analysis was performed which consisted firstly of calculating the frequencies of each type of sequence and secondly of the rate of the occurrence of problematic answers. This analysis provided information about questions with possible difficulties. Then, an evaluation of the convergence was done to obtain information about the agreement in the answers given by both types of informant. To carry out the convergence analysis, the final response of every self-reporter in each of the questions was compared with the final response given by his proxy in the same question.

3 Results

For the analysis 352 sequences (i.e., 11 questions \times 32 respondents) were taken into account.

3.1 Sequences types analysis

First, the sequence types produced by both types of informant were compared. Table 4 shows the results from this comparison.

As Table 4 shows, significant differences ($\chi^2 = 15.706$; $p < 0.001$) were found in the percentages of the types of sequence produced by both types of informant. The greater differences occur in the percentage of paradigmatic sequences, these being higher for proxy reporters. Also, self-reporters show a high percentage of non paradigmatic-problematic sequences. Thus, self-reporters not only deviate from the paradigmatic pattern more often than proxy-reporters, but also these deviations are more often problematic.

Table 4 Percentage of different sequences produced by "proxies" and self-reporters

Sequence type	Self-reporters	Proxies reporter
Paradigmatic sequences	38	58
Non paradigmatic-non problematic	23	19
Non paradigmatic-problematic	39	23

$$\chi^2 = 15.706 \quad p < 0.001$$

Table 5 Percentages of problematic and non problematic answers

Question number	“Any problematic answer”		“Adequate answer”	
	Self	Proxy	Self	Proxy
Q.1. “Habitual activities”	44	75	81	88
Q.2. “Speak”	44	19	69	81
Q.3. “Understand”	56	6	56	100
Q.4. “Use the phone”	69	44	81	75
Q.5. “Use the senses”	31	31	88	94
Q.6. “Learn”	38	25	94	100
Q.7. “Move the body”	31	19	94	100
Q.8. “Change posture”	44	19	63	94
Q.9. “Show affection”	31	13	88	94
Q.10. “Family relationships”	19	0	81	100
Q.11. “Sentimental relationships”	25	19	88	94

3.2 Comparison between codes produced for proxies and self-reporters

Next, the frequency of the answering behaviours was analyzed for both types of informant. First, the types of answer given by the respondent were observed. The answers were classified in two groups: problematic answers, where answers coded as mismatch answer, invalid answer, don’t know answer and qualified answer were included; and non problematic answers composed by answers coded as adequate. The last one includes the sequences in which an adequate answer was given although other problematic answers occurred beforehand. Because of the existence of multiple behaviours in the same sequence, the total percentage can exceed 100%. Table 5 shows the percentage of problematic and non problematic answers to each question of the questionnaire for both types of informant.

As Table 5 shows, in general, self-reporters show higher percentages of problematic answers than proxies except in question 5 “use the senses” where the percentage is equal for both (31%) and in question 1 “habitual activities” where proxies produce problematic answers in a very high percentage (75%). This last question and question 4 “use the phone” reached the highest percentages of problematic answers for both informants. Also, question 1 “habitual activities”, question 3 “understand” and question 10 “family relationships” show the highest differences between both informants. In the first case, proxies gave more problematic answers than self-reporters, while in question 3 and question 10 self-reporters produced more problematic answers than proxies, who never produced a problematic answer in question 10. On the other hand, the percentage of non problematic answers, that is to say, of adequate answers was always greater for proxies except in question 4 “use the phone”. The largest differences between both informants occurred in question 3 “understand” where proxies reached 100% of adequate answers.

In addition, the percentage of behaviours occurring while the interviewer was asking the question was analyzed for both types of informant. Table 6 shows the percentages of the occurrence of these additional behaviours.

As Table 6 shows, the percentages obtained for the code ‘request clarification’ are in general higher for self-reporters than for proxies, except for the questions 2 “Speak” and 10 “Family relationships”. A striking difference is that in the questions 6 “Learn” and 7 “Move the body” the percentages of occurrence are high for self-reporters whereas proxies never

Table 6 Percentages of the behaviours occurring while asking the question

Question	Code			
	Request clarification		Interruption	
	Self	Proxy	Self	Proxy
Q.1. "Habitual activities"	6	6	6	0
Q.2. "Speak"	0	6	0	6
Q.3. "Understand"	6	0	0	6
Q.4. "Use the phone"	13	13	0	0
Q.5. "Use the senses"	6	0	0	0
Q.6. "Learn"	19	0	0	0
Q.7. "Move the body"	19	0	0	0
Q.8. "Change posture"	6	6	0	0
Q.9. "Show affection"	6	0	0	0
Q.10. "Family relationships"	0	13	0	0
Q.11. "Sentimental relationships"	13	0	0	0

produce requests for clarification in these questions. As for the interruptions, they appear with higher frequency for proxies than for self-reporters, though the percentages are not high.

3.3 Convergence evaluation

After the informants' behaviour analysis, the convergence between both types of informant was evaluated. Two approaches can be used to compute the disagreement between both types of informant. The "traditional" approach calculates the percentage of disagreement taking all sequences into account. In doing that, the percentage of disagreement was 48% and the percentage of sequences with agreement in the answers given for both members of the same couple was 52%.

The "traditional" approach uses all sequences no matter if the sequences are "problematic" or "non problematic". When that approach is used, researchers miss that the convergence evaluation is not always possible or, at least, advisable. There are, for instance, situations in which one member of the couple did not give an answer or the answer given was not an adequate answer. For example, if in a yes/no question the proxy says "yes" and the self-reporter says "no" there is disagreement, but if proxy says "yes" and the self-reporter says "sometimes", there is a mismatch answer, that is to say an answer which does not fit to any of the alternatives given. Traditionally, these situations have been considered as "disagreement situations".

The behaviour coding method allows us to optimize the convergence evaluation by only selecting cases in which both proxy and self-reporters gave an adequate answer. This selection is important for knowing the real percentage of disagreement and to filter the evaluation by removing situations in which a final adequate answer was not obtained. After removing these cases, the percentages of disagreement were reduced to 19%.

Furthermore, the "disagreement in answers" sequences found when either proxy or self-reporters give a non adequate answer, can be categorized into three groups: (a) "Self non adequate", when the self reporter did not give an adequate answer but the proxy did; (b) "Proxy non adequate", when the contrary occurred; and (c) "Both non adequate", when

Table 7 Percentage of sequence with agreement and disagreement

Agreement/disagreement	Percentage
Adequate answers	
Agreement in answers	52
Disagreement in answers	19
Non adequate answers	
Self non adequate	23
Proxy non adequate	6
Both non adequate	0

Table 8 Agreement between proxy and self-reporter answers

Agreement/disagreement	Questions numbers										
	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10	Q.11
Agreement in answers	38	38	31	44	63	75	56	38	56	56	75
Disagreement in answers	19	13	19	13	19	19	31	19	25	25	6
Self non adequate	25	38	50	25	13	6	13	38	13	19	19
Proxy non adequate	19	13	0	19	6	0	0	6	6	0	0

neither of them gave a final adequate answer. Table 7 shows the percentages for each of these categories together with the percentages of agreement and disagreement when both types of informant ended up giving an adequate answer.

Table 7 shows how the percentage of disagreement in the answers descends to 19% due to the percentage of non adequate answers (29% in total) mainly on the part of the self-reporter (23%). To carry out the convergence analyses between the answers provided by self-reporters and proxies, only the percentages of adequate answers were included.

In addition, the percentages of agreement and disagreement throughout the set of target questions were calculated. Table 8 presents the percentages of agreement/disagreement in answers calculated by only counting final “adequate” answers. Table 8 also shows the percentages of non adequate answers in order to detect cases with high differences between both types of informant.

Table 8 shows differences across the different questions. A high percentage of disagreement exists for question 7 “Family relationships” (31%), while questions 6 “Learn” and 11 “sentimental relationships” achieve the highest degree of agreement (75%). In relation to non adequate answers, the largest percentages are observed in self-reporters for all the questions. For example, self-reporters have the highest percentage of non adequate answers (50%), while the percentage achieved by the proxies in the same question was 0.

4 Discussion

The objectives of the study were to analyze the behaviour of proxy reporters and to evaluate the convergence between the answers given by self-reporters and proxies to a disability questionnaire by means of a behaviour coding method. The rationale behind applying a behaviour coding method was to improve the evaluation of convergence by analyzing only

final adequate answers i.e., answers that are directly interpretable as an answer, given by both types of informant.

The analysis of the behaviour of both types of informant showed that the percentage of paradigmatic sequences in both groups can be considered adequate. [Van der Zouwen and Smit \(2004\)](#) found in a study 29.8% of paradigmatic sequences. However, this percentage was higher in the group of proxy reporters (37.5 vs. 57.95%). The opposite difference was found with regard to the percentages of non paradigmatic-problematic sequence (39.20 vs. 23.30%). The percentage of non paradigmatic-problematic sequences is slightly higher than the percentages found in the bibliography. [Dijkstra and Ongena \(2006\)](#) found percentages between 22 and 37.9% of problematic sequences in the analysis of five different surveys.

The extent to which both types of informant perform the stages of the cognitive “question-and-answer” process may explain the differences in the percentages for the types of sequence. According to the Krosnick’s theory of “optimizing vs. satisficing”, the involvement of the respondents when answering survey questions depends on three points: task difficulty, respondents’ ability and respondents’ motivation ([Krosnick 1999](#)). Based on the findings of behaviour coding in this study, the difficulty of the task and the motivation for responding to questions could be different for both types of informant. Self-reporters may have more information available in the episodic memory, which, in contrast to first impressions, could make it more difficult to translate to a response category. Among the aspects which may influence the participants’ motivation, the personal importance of the question’s topic to the respondents could be the most important. In this study, the questionnaire topic might be more important for the self-reporter because they report on their own situation. On the other hand, proxies report on the situation of another person. All these aspect taken together could increase the chance of a high number of problematic answers for self-reporters, increasing the probability of non paradigmatic sequences.

In the specific analysis of the appearance of problematic and non problematic answers, self-reporters showed a higher percentage of problematic answers than proxies. In this analysis, it is important to point out three main results. Firstly, questions with a high percentage of problematic answers for both informants, such as question 4 “use the phone”, might indicate possible problems with the wording or the format of the question. Question 4 has a complex format because it includes instructions about some information that the respondent should exclude while answering, which is a challenging cognitive task. Secondly, in questions in which proxies have obtained higher percentages of problematic answers, such as question 1 about limitations to habitual activities, the percentage of problematic answers might be due to the proxies’ lack of detailed information about the topic. Finally, there are questions in which self-reporter obtained higher percentages of problematic answers, such as question 3 and 10.

These results might be explained by the optimizing theory. According to the optimizing theory a respondent who is optimizing would carefully assess the appropriateness of each response before selecting one. In contrast, a respondent who is satisficing could simply choose the first reasonable response ([Krosnick 1999](#)). Considering that self-reporters are optimizing, because the topic is more important for them and they should be more motivated, it is possible they analysed the alternatives, which are “yes” and “no” (all questions are yes/no questions) determining that neither of them were completely adequate. In this case, the self-reporter probably gave an invalid or mismatch answer. On the other hand, proxies are satisficing and they maybe selecting the closest alternative to the real situation, and for this reason they achieved 100% of adequate answers, but which may not necessarily be the most correct answer.

The convergence analysis showed that the percentage of agreement was 52% and how the percentage of disagreement falls from 48 to 19% when counting only final adequate answers. Both results were obtained by analyzing “comparable” answers given by proxies and self-reporters. On the other hand, the convergence analysis for each of the target questions showed higher percentages of agreement in questions 6 and 11, whereas question 7 reaches the highest percentage of disagreement. The content of question 6 was about difficulties when writing, copying, counting or using everyday utensils. As [Magaziner et al. \(1996\)](#) point out, proxies inform accurately about physical or daily tasks, and this might be the reason for the high agreement found. Question 11 is about sentimental and sexual relations. In this case, the intimate content of the question could cause both informants to become satisficers, influenced by social desirability, thereby reaching a high percentage of agreement. In relation to disagreement situations, question 7 obtained the highest percentage. This question had a complex format and its content was also complex. It could be possible both informants went through a different satisficing process (i.e., strong or weak satisficing) to answer this question. As Schwarz and Wellens have already pointed out, the information retrieval and judgment process is likely to proceed differently for proxy vs. self-reporters.

In conclusion, in spite of the results found by [Loftus et al. \(1992\)](#) and [Todorov and Kirchner \(2000\)](#) in their studies with proxies and self-reporters, the results show a better behaviour of the proxy reporters from an interviewer-respondent interaction perspective. However, although the answers given by proxies have been adequate in a high percentage and the behaviour while asking the question has been less problematic, we have to take into account that satisficing can not be detected as clearly by means of behaviour coding. Thus, it is possible that although they gave adequate answers, the proxies were satisficing. For this reason, although it is possible to say the convergence between proxies and self-reporters exists because a high percentage of agreement has been obtained, it is necessary to have more detailed information on the cognitive process taken by both informants to assure that their answers are really equally valid. Thus future research could be focused on applying other procedures, such as cognitive interviews, to obtain more information.

In relation to the usefulness of behaviour coding, this procedure has made two fundamental contributions. First, it enables analyses with the information obtained directly from the interaction, i.e., what has actually happened during the interview. Traditionally, this type of analysis has been carried out using the information registered by the interviewer in the questionnaire, which could be more biased than the information proceeding directly from the record of the interaction. Second, behaviour coding has allowed selecting the analysis only for those cases in which both informants had given an adequate answer. Including only these cases yields a truly fair comparison of proxies and self-reporters, since the interactional situation of the answer was taken into account.

References

- DeMaio, T.J., Rothgeb, J., Hess, J.: Improving survey quality through pretesting. U.S. Bureau of the Census, Washington. http://www.amstat.org/Sections/Srms/Proceedings/papers/1998_007.pdf (1998). Accessed 30 July 2009
- Dijkstra, W.: Sequence Viewer (version 4.4a). Free University of Amsterdam, Netherlands (2008)
- Dijkstra, W., Ongena, Y.: Question-answer sequences in survey-interviews. *Qual. Quant.* **40**(6), 983–1011 (2006)
- Duncan, P., Min Lai, S., Tyler, D., Perera, S., Reker, D.M., Studenski, S.: Evaluation of proxy responses to the stroke impact scale. *Stroke* **33**, 2593–2599 (2002)

- Forsyth, B., Levin, K., Fisher, S.: Test of an appraisal method for establishment survey questionnaires. In: *Proceeding of the ASA Section on Survey Research Methods*. American Statistical Association, Alexandria (1999)
- Krosnick, J.A.: Survey research. *Annu. Rev. Psychol.* **50**, 537–567 (1999)
- Loftus, E.F., Smith, K.D., Klinger, M.R., Fiedler, J.: Memory and mismemory for health events. In: Tanur, J.M. (ed.) *Questions About Questions: Inquiries into the Cognitive Bases of Surveys*, pp. 102–137. Russell Sage, New York (1992)
- Magaziner, J., Speaar, S., Hebel, J.R., Gruber-Baldini, A.: Use of “proxies” to measure health and functional status in epidemiologic studies of community-dwelling women. *Am. J. Epidemiol.* **143**(3), 283–292 (1996)
- Oksenberg, L., Cannell, C., Kalton, G.: New strategies for pretesting survey questions. *J. Off. Stat.* **7**(3), 349–365 (1991)
- Ongena, Y.P.: Interviewer and respondent interaction in survey interviews. Unpublished doctoral dissertation. Amsterdam Vrije Universiteit (2005)
- Ongena, Y.P., Dijkstra, W.: Methods of behavior coding of survey interviews. *J. Off. Stat.* **22**, 419–451 (2006)
- Pickard, A.S., Johnson, J.A., Feeny, D.H., Ashfaq, M.D., Carriere, K.C., Abdul, M.N.: Agreement between patient and proxy assessments of health-related quality of life after stroke using the EQ-5D and health utilities index. *Stroke* **35**, 607–612 (2004)
- Schaeffer, N.C., Maynard, D.W.: From paradigm to prototype and back again: interactive aspects of cognitive processing in standardized survey interviews. In: Schwarz, N., Sudman, S. (eds.) *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research*, pp. 367–391. Jossey-Bass, San Francisco (1996)
- Schwarz, N., Wellens, T.: Cognitive dynamics of proxy responding: the diverging perspectives of actors and observers. *J. Off. Stat.* **13**(2), 159–174 (1997)
- Sunghee, L., Mathiowetz, N.A., Tourangeau, R.: Measuring disabilities in surveys: consistency over the time and across respondents. *J. Off. Stat.* **23**(2), 163–184 (2007)
- Tafforeau, J., Lopez, M., Tolonen, A., Scheidt-Nave, C., Tinto, A.: European Commission. Eurostat. Guidelines for the development and criteria for the adoption of Health Survey instruments: <http://ec.europa.eu/health/ph_information/dissemination/reporting/healthsurveys_en.pdf> [Check: March 22, 2011] (2006)
- Todorov, A., Kirchner, C.: Bias in “proxies” reports of disability: data from the national health interview survey on disability. *Am. J. Public Health* **90**(8), 1248–1253 (2000)
- Tourangeau, R.: Cognitive aspects of survey measurement and mismeasurement. *Int. J. Public Opin. Res.* **15**, 3–7 (2003)
- Tourangeau, R., Rips, L.J., Rasinski, K.: *The Psychology of Survey Response*. Cambridge University Press, Cambridge (2000)
- Vander Zouwen, J., Smit, J.H.: Evaluating survey questions by analyzing patterns of behaviour codes and transcripts of question-answer sequences: a diagnostic approach. In: Presser, S., Rothgeb, J., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Singer, E. (eds.) *Methods for Testing and Evaluating Survey Questionnaires*, pp. 109–130. Wiley, New York (2004)
- World Health Organization: ICFH-2: International Classification of Functioning, Disability and Health (ICF). Author, Geneva (2009)