Statistica & Applicazioni - Vol. XIV, n. 1, 2016, pp. 29-48

USING IRT MODELS TO QUANTIFY THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRE (SDQ) OUTCOMES

Simone Di Zio*
Giulia di Francesco**
Annalina Sarra**

SUMMARY

Over the past decades an increasing number of studies have focused on the importance of behavioural problems of school children. Often, the assessment of children's behavioural and emotional problems has been carried out using the Strengths and Difficulties Questionnaire (SDQ). This paper analyses the parents' and teachers' scores of SDQ for a sample of children aged 6 to 10 years who participated in a karate project in a public school. To handle the response options of the SDQ we relied on Item Response Theory (IRT) models. In particular, in this setting, we exploited the attractive features of the Linear Logistic Models with Relaxed Assumptions to measure the change in the SDQ dimensions parameter estimates which occurred over two measurement occasions: before and after the karate project. Notwithstanding the wide and ongoing use of the SDQ as brief behavioural screening there are no studies, to date, which have illustrated the usefulness of this class of IRT-based model for assessing change in the SDQ dimensions parameter estimates over time. This paper aims to fill this gap and discusses the main results of the application.

Keywords: *Strengths and Difficulties Questionnaire, Item Response Theory Models, Linear Logistic Models with Relaxed Assumptions, Measurement of Change, Emotional and Behavioural Problems.*

1. INTRODUCTION

The paper concerns an application of the Strengths and Difficulties Questionnaire (SDQ) on a sample of children between 6 and 10 years, who participated in a school project of karate in a village in the province of Pescara (Italy). The project lasted from January to June 2012 and consisted in one hour of physical education per week based on the principles of eastern martial arts (Di Zio, 2009).

The aim of this study is to exploit the potentialities of the Item Response Theory (IRT) models for the analysis of the Strengths and Difficulties Questionnaire scores.

The SDQ is a 25 items tool designed to collect data on different psychological aspects of children and adolescents (Goodman, 1997; Goodman, Meltzer and Bailey, 1998).

* Dipartimento di Scienze Giuridiche e Sociali - Università ''G. d'Annunzio'' - Viale Pindaro, 42 - 65127 PESCARA (e-mail: s.dizio@unich.it).
** Dipartimento di Economia - Università ''G. d'Annunzio'' - Viale Pindaro, 42 - 65127 PESCARA (e-mail: ✉ giulia.difrancesco@unich.it; asarra@dmqte.unich.it).

The SDQ can be completed by parents and teachers to evaluate children's behaviour. Note that, there is an initial version and a follow-up questionnaire for use after an intervention (i.e. an educational program, a daily clinical practice, etc.), which includes two additional questions.

IRT models are a family of measurement techniques concerned with the measurement of an individual's latent traits (such as attitude, ability, skill, satisfaction) assessed indirectly by a group of items (De Mars, 2010). These models, starting from raw scores, are able to construct continuous, linear measures of both items and subjects. Through the IRT models, the subject and item parameters are expressed according to a common measurement unit on the same continuum, becoming directly comparable. Accordingly, the IRT models allow to examine latent constructs in a rigorous way, providing more accurate and consistent results compared to traditional measurement instruments, such as Classical Test Theory (de Ayala, 2009).

Several authors have introduced and discussed the advantages of applying IRT models in behavioural and social sciences (e.g., Embretson and Reise, 2000) to construct and validate measurement scales as well as to model and interpret items response data.

But, despite the large body of literature on IRT research in general, only a few studies have directly applied the IRT models to measure the outcomes of the SDQ. As far as we know, previous researches based on IRT models, and in particular on Rasch models, have focused on assessing the psychometric properties of this survey instrument (see, for instance, Hagquist, 2007).

Conversely, in many studies dealing with SDQ data (Janssens and Deboutte, 2009; Giannakopoulos, Tzavara, Dimitrakaki, Kolaitis, Rotsika, and Tountas, 2009; Goodman, Lamping and Ploubidis, 2010), the Factor Analysis (FA) is traditionally applied (Stokes, Mellor, Yeow and Hapidzal, 2014; Di Riso, Salcuni, Chessa, Raudino, Lis and Alto, 2010). Even if the FA and the IRT have ''virtually identical statistical formulation'' (Reckase, 2009), there are many differences between the two approaches and the IRT models offer some additional benefits.

The IRT models, unlike other methods allow evaluation at the same time of both the characteristics of the subjects and the characteristics of the items (De Boeck and Wilson, 2004).

The varying characteristics of items such as the difficulty, the discrimination power and the possibility of guessing (i.e. the possibility that a person might guess the correct answer) are of fundamental importance in the IRT while in the FA are not considered as parameters to be estimated, but only as nuisance to be removed (Reckase, 2009). Furthermore, while the FA is mainly a data reduction technique, the IRT can model the interaction between persons and test items. Importantly, the IRT models, using the same latent space for tests and samples, offer opportunities to express the subject and item parameters according to a common measurement unit on the same continuum.

Differently from the most of the studies dealing with SDQ, which concentrates only on the initial or final version of the questionnaire, in this paper we propose an

attempt to jointly estimate the person and item parameters of the two versions of this instrument.

In particular, by using the Linear Logistic models with Relaxed Assumption (LLRA), we are able to estimate the amount of changes during the elapsed time between two subsequent administration of the test (Hatzinger and Rusch, 2009; Mair and Hatzinger, 2008). Furthermore, by this approach it is also possible to measure the SDQ outcomes and investigate if there are any differences between teachers and parents.

Besides, since the original five factor structure of SDQ is not always supported by empirical researches, in this study we acknowledge the need for a different SDQ formulation. In this respect, we compare two multidimensional IRT models, which account for a different internal structure of SDQ and we rely on the usual goodness-of-fit statistics as a guide for our model selection.

The remainder of the paper proceeds as follows. In Section 2 we give main details on the measurement instrument (SDQ). Section 3 is devoted to present the karate project while Section 4 deals with the statistical background of IRT models. In particular, linear logistic models with relaxed assumptions, aimed at capturing change in parameter estimates over time, are illustrated. Results and concluding remarks can be found in Sections 5 and 6, respectively.

2. THE STRENGTHS AND DIFFICULTIES QUESTIONNAIRE

The SDQ is a brief psychopathology screening tool administered to parents and/or teachers to assess behavioural and emotional problems in children and adolescents (Goodman, 1997; Goodman *et al*., 1998).

The items included in the SDQ refer to the positive or negative attributes of a child's behaviour and are grouped in five subscales relating to emotional problems, conduct problems, hyperactive-inattention, peer relationship problems, and prosocial behaviour. Each subscale consists of five items with 3-point response scale (Not true = 0, Somewhat true = 1, Certainly true = 2). Positively and negatively worded items were interchanged to avoid bias. More specifically, for the items 7, 14, 15, 17 and 18, which express positive qualities, it has been necessary to reverse the scale before making any statistical analysis.

The sum of four of the five subscale scores (the prosocial scale is excluded) yields a total difficulties score.

The internal structure of the SDQ was developed with reference to the main nosological categories recognised by contemporary classification systems of child mental disorders such as the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV; American Psychiatric Association, 1994).

The SDQ has been extensively validated. In particular, a number of studies, mainly in Europe, have provided consistent support for the hypothesised five-subscales through explanatory factor analysis (Achenbach, Becker, Dopfner, Heiervang, Roessner, Steinhausen and Rothenberger, 2008; Becker, Steinhausen, Baldursson, Dalsgaard, Lorenzo, Ralston, Döpfnex and Rothenberger, 2006; Smedje, Broman,

Hetta and von Knorring, 1999; Woerner, Fleitlich-Bilyk, Martinussen, Fletcher, Cucchiaro Dalgalarrondo, Lui and Tannock, 2004). Other scholars relied on a model-based framework, such as confirmatory factor analysis (CFA), to assess the original five-factors structure of the SDQ (see, among others, He, Burstein, Schmitz and Merikangas, 2013).

As will be detailed in the next sections, in this study, in line with a confirmatory perspective, the number of SDQ latent dimensions and the relationship with items are specified in advance. Thus, the focus will be concerned with the evaluation of the departure from these prior specifications and the comparison with alternative factor structures, justifiable on theoretical grounds.

The SDQ is widely used by teachers, educators and other professionals, to identify children at risk of psychological disorders (Goodman, Ford, Simmons, Gatward and Meltzer, 2000). Early recognition of psychological disturbances is of fundamental importance since these problems may affect the individual psychological well-being as well as the child's future development.

As said before, the SDQ consists of two versions: initial and follow-up. By administering the follow up version after an intervention it is possible to understand if the problems are reduced or if the procedure has helped in making the problems more acceptable. Compared to the initial version, the follow-up of SDQ focuses on a shorter time-frame, to investigate the problems relevant to behaviour of children and adolescent, increasing the possibility of detecting changes in children's behaviour.

The SDQ is available in over 30 languages and is extensively applied throughout the world, including numerous European countries (United Kingdom, Finland, Germany, Sweden and Italy). In this paper, we used the Italian version of the SDQ, already employed in other researches to reduce children's behavioural problems at school or assess the prevalence of mental disorders in preadolescents (Marzocchi, Capron, Di Pietro, Duran Tauleria, Duyme, Frigerio, Gaspar, Hamilton, Pithon, Simoes and Therond, 2004).

Several works (see, among others, Goodman, 2001; Hawes and Dadds, 2004) confirm the good psychometric properties of the instrument and the excellent relationship between time taken to complete it and the amount of information collected.

The SDQ has been used in various ways, in particular in clinical assessment, but there are also applications in epidemiological studies, genetic, social studies and educational settings (Goodman, 1997). For instance, the instrument has been applied with success in the diagnosis of Attention Deficit-Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder (ODD) and Conduct Disorder (CD), which according to the DSM IV are included in the category of Disruptive Behaviour Disorder (DBD) (Ullebo, Posserud, Heiervang, Gillberg and Obel, 2011).

3. THE PROJECT OF MARTIAL ARTS IN AN ITALIAN PRIMARY SCHOOL

We have used the SDQ as part of a project of physical activity and martial arts in an Italian primary school, in 2012, in the village of Rosciano, in the Abruzzo region

(Italy). Three expert Karate masters conducted a lesson of one hour a week, for each class, in the school gym during the curricular hours. The project involved all the 142 pupils of the school, aged 6 to 10 years, belonging to 8 different classes. Just before starting the project, the initial version of the SDQ questionnaire was completed by parents and teachers, and during the last week of June 2012 it was administered the follow-up version.

In the first phase, we collected 117 questionnaires, completed by parents and 142 by teachers; in the second stage of the project (administration of the follow-up version of the SDQ) 127 questionnaire were compiled by parents and 142 from teachers, for a total of 528 valid questionnaires. Therefore, while the response rate of teachers was 100% in both administrations, confirming their full cooperation to the study, an overall rate of response of 82% and 89% was obtained for the parents, in the first and second phase, respectively. Note that the project lasted less than the entire school year, because for administrative reasons it was not possible to start it at the beginning of the school year.

The karate project was developed by the Italian Federation of Judo, Fight, Karate, and Martial Arts (FIJLKAM) which is among the first sport federations that have obtained the approval of a project for teaching its disciplines in public schools. Through a large number of clubs of martial arts, scattered throughout the country, the federation has entered primary and secondary schools, since the early 2000s.

A martial art, revisited with the critical methods of modern physiology, biomechanics and pedagogy, is a useful tool to achieve the purposes of the modern physical education (Fabbri, Fabbri and Primi, 2001; Saibene, Rossi and Cortili, 1995; Aschieri, 2005). The main goal is to form skills that go beyond the physical context, in order to develop abilities in performing cognitive operations, in a timely, effective and creative way (Gardner, 1983). In addition, the project is aimed at training persons who can interact with each other and with the environment, following a scheme based on precise rules. It is well acknowledged that the ability to interact with a partner, regardless of race, religion and gender, is very important for the improvement of the self-awareness and self-control and for the 'respectful mind' (Gardner, 1983).

The potential value of the introduction of martial arts in physical education programs has been documented in many researches (Banks and Reed, 2003; Brown and Johnson, 2000; Lakes and Hoyt, 2004; Reilly and Friesen, 2001; Theeboom and De Knop, 1999), which also stress the positive effect of martial arts participation also on school performance (Vockell and Kwak, 1990).

Accordingly, the team of experts who has created the project believes that the practice of physical activities in primary school, based on the principles of eastern martial arts, may have positive impact in the spheres of children's behaviour such as emotional symptoms, hyperactivity/inattention and prosocial behaviour (Aschieri, 2005; Nosanchuk, 1981; Walsh, 2000).

4. METHODOLOGICAL FRAMEWORK: IRT MODELS

The IRT encompasses a variety of statistical methods to assess the consistency between a latent trait of interest and the specific responses to a set of items, differing in their mathematical form and underlying assumptions. All IRT models include a latent trait parameter and an item location parameter and link the probability of a subject $p(p = 1, ..., P)$ providing a certain answer (Y_{pi}) to an item $i(i = 1, ..., I)$ given the individual's latent trait and one or more item characteristics.

According to the special case of IRT models for binary data $(Y_{pi} = \{0, 1\})$, the probability of observing a given response is modelled as a function of the difference between the latent trait (θ_p) expressed by the *p-th* subject, and the difficulty (β_i) on the *i-th* component (item) of the test.

The basic IRT model for dichotomous data (and a single dimension), also known as the one-parameter Logistic Model (De Boeck and Wilson, 2004), or Rasch Model, can be expressed as:

$$\Pr\left(Y_{pi} = 1 \middle| \theta_p, \beta_i\right) = \frac{\exp\left(\theta_p - \beta_i\right)}{1 + \exp\left(\theta_p - \beta_i\right)} \tag{1}$$

An IRT model for dichotomous data operates under the assumption that the items share a common underlying construct (*unidimensionality*), i.e. it is supposed that the item responses are explained by one latent trait, and *local independence*. This last condition implies that, given the value of the latent trait, the item responses are distributed independently. In fact, the classical formulation of IRT models typically does not consider a nested structure of data. To simultaneously estimate item and person parameters and naturally accommodate the hierarchical structure of the data, a multilevel IRT modelling can be adopted (Maier, 2001; Kamata, 2001; Beretvas and Kamata, 2005). The multilevel IRT models are advantageous in that they incorporate random effects to account for various dependencies found in the data (Fox, 2007).

In addition, in real applications, the interactions between persons and test items are not as simple as implied by unidimensional models, because generally the problems posed by test items require numerous skills and abilities to determine a correct solution. Therefore, there is the need for more sophisticated IRT models that reflect the complexity of the interactions between examinees and test items (Reckase, 2009). Many psychological constructs are, unavoidably, multi-dimensional in nature, namely an item or a set of items measure more than one latent trait. Latent traits might be understood as a combination of sub-scale components, nested within a more general construct (Reckase, 2009).

An extension of uniunidimensional IRT models, suitable for describing situations where multiple skills are needed to respond to test items, is that of Multidimensional Item Response Theory (MIRT) models (Reckase, 2009). These models have multiple parameters for the person, reflecting the hypothesis that persons vary in a wide range of traits. Technically, MIRT turns out to be a special case of IRT, since in addition to a vector of item characteristics describing the item test difficulty there is also a

vector of multiple person characteristics that describe abilities that a subject expresses in responding to a test. The multidimensional model is defined as:

$$\Pr\left(Y_{pi}=1|\theta_{pd},\beta\right)=\frac{\exp\left(\theta_{pd}-\beta_{id}\right)}{1+\exp\left(\theta_{pd}-\beta_{id}\right)} \tag{2}$$

where d is an index for the dimension ($d=1,2,...,D$), β_{id} is the difficulty parameter for dimension d and item i, while θ_{pd} is the ability along dimension d (D is the number of dimensions in the model).

Literature proposes different variants of MIRT models, and an important class of MIRT model can be defined assuming a different complexity of the statistical relation between dimensions and items. Following a widespread terminology (Adams, Wilson and Wang, 1997), it is possible to distinguish models with between-item multidimensionality from models with within-item multidimensionality. In models that incorporate between-item multidimensionality, each item only loads on one dimension; conversely models including within-item multidimensionality involve a more complex loading structure, where each item may potentially load on all the latent traits. In this work, we adopt the former formulation which requires that each subset of items measures only one latent trait.

The restrictive assumption of unidimensionality also drops in the LLRA, which is a class of IRT models specifically developed to estimate the amount of changes during the elapsed time between two subsequent administration of the same test (Hatzinger and Rusch, 2009; Mair and Hatzinger, 2008). The LLRA models, first introduced by Fisher (1974), are generalised Rasch models with multidimensional latent trait parameters in which changes are modelled as a function of treatment effects, treatment interactions, and trend effects (Hatzinger and Rusch, 2009).

Given two different time points (T_1 and T_2), the LLRA model for dichotomous data can be formalised in the following equations:

$$\Pr\left(Y_{pi1}=1\middle|T_1,\theta_{pi},\beta_i\right)=\frac{\exp\left(\theta_{pi}-\beta_i\right)}{1+\exp\left(\theta_{pi}-\beta_i\right)} \tag{3}$$

$$\Pr\left(Y_{pi2}=1\middle|T_2,\theta^*_{pi},\beta_i\right)=\frac{\exp\left(\theta_{pi}+\delta_{pi}-\beta_i\right)}{1+\exp\left(\theta_{pi}+\delta_{pi}-\beta_i\right)}=\frac{\exp\left(\theta^*_{pi}-\beta_i\right)}{1+\exp\left(\theta^*_{pi}-\beta_i\right)} \tag{4}$$

where $\Pr\left(Y_{pi1}=1\middle|T_1,\theta_{pi},\beta_i\right)$ and $\Pr\left(Y_{pi2}=1\middle|T_2,\theta^*_{pi},\beta_i\right)$ are the probabilities for subject p to score 1 on item i at T_1 and T_2, respectively, while θ_{pi} (the latent trait) is the location of subject p on the *i-th* item at T_1 and $\delta_{pi}=\theta^*_{pi}-\theta_{pi}$ is the quantity of change, on item i, for the *p-th* subject.

Usually the quantity δ_{pi} is modelled as the sum of group-specific (as for instance a treatment) effect parameters and unspecified trend parameter (changes independent of the treatment, such as retest effects).

Following Hatzinger and Rusch (2009) the quantity δ_{pi} can be explicited as:

$$\delta_{pi} = \sum_{j} q_{pji}\lambda_{ji} + \tau_i + \sum_{j<l} q_{pji}q_{pli}\rho_{jli} \tag{5}$$

In (5) q_{pji} stands for the dosage of treatments j for item i in subject p, λ_{ji} denotes the effect of the treatment j on item i, τ_i is the parameter for the trend effect on item i, between T_1 and T_2, and ρ_{jli} are the parameters for interaction effects of treatments j and l on item i (Hatzinger and Rusch, 2009). The linear re-parametrization of δ_{pi}, as shown in (5), assures to the LLRA models a great flexibility, allowing to give valid responses to different research questions. More specifically, through these models one can test if there is a trend effect and/or a treatment/covariate effect as well as if the treatment/covariate/trend effects are the same for a group and if there are any interaction effects between groups (e.g. gender and treatment). Notice that the model specified in the previous equations is the most general one in terms of dimensions because it is assumed that each item measures a single latent trait. If desired it is possible to account for groups of items which measure the same latent trait.

In our study, we will focus on the estimation of τ_d, for $d = 1, ..., m$ and $m = 5$ representing the five dimensions of the SDQ questionnaire so that we can obtain estimates of the changes in the five dimensions before and after the karate project, along with separate estimates of τ_d for parents and teachers. In particular, negative values of τ_d indicate an improvement in dimension d while positive values denote a worsening.

A detailed presentation of LLRA models can be found in Fischer (1974) and Fischer and Ponocny (1995) where their useful properties for the measurement of change and their estimation are highlighted. Here, it is sufficient to say that the parameter estimation in a LLRA model is based on conditional maximum likelihood estimation.

5. MAIN RESULTS OF IRT ANALYSIS

In the previous section, it was discussed how IRT models and the class of LLRA models can provide a suitable approach for handling the outcomes arising from the SDQ.

In particular, the attractive features of LLRA model, which does not require unidimensionality of the items or distributional assumptions about the population of persons, are of interest in our context to measure the changes in the scores of SDQ scales over time.

The initial three response categories for all 25 SDQ-items, which are in the form of statements, have been dichotomized by aggregating the two categories ''somewhat true'' and ''certainly true'' in a unique category (say 1) and leaving the value 0 with the meaning of ''not true''. As stated before, the response categories are ordered in terms of implied agreement to the statements and the items scored in the opposite direction were recoded. In terms of the IRT jargon, in our context, the ability parameter θ_p assumes the meaning of ''level of behavioural difficulty'', so that the higher the

value of theta, the greater the behavioural problem of the child. From the item side, the difficulty parameter β_i should be interpreted as ''severity of the item'', in the sense that the higher the value of beta, the greater the difficulty of declaring that the child behaves as stated by the item *i-th*. For example, the item ''steals from home, school or elsewhere'' is certainly more severe than the item ''Easily distracted''. In other terms, the probability that a teacher, or a parent, gives a positive answer to the first item is definitely lower than the probability of a positive response to the second one.

In what follows, we first provide the results of a descriptive IRT model (one parameter logistic model), defined assuming a two-level structure in the data: item responses as the level one variable, nested within subjects.

We specify a model in which the inherent multilevel structure of the data (the clustering of the item responses within respondents) is a function of item-specific fixed effects and one-person specific random components.

The results were obtained using the R package lme4 (Bates, Maechler and Bolker, 2013) and refer to the entire data set, including responses of teachers, parents to both questionnaires (initial and follows up versions). A total of 528 valid questionnaires form the basis of this preliminary data analysis. The fixed effect parameter estimates ($\hat{\beta}_i$) for this model are displayed in Table 1, along with the associated standard errors and their statistical significance.

Upon inspection of Table 1, it is evident that only the item 14 (*Thinks things out before acting*) is not statistically significant.

It is important to recall that the smaller values of these estimates describe ''the easiest items'', that is, the statements that many respondents are likely to endorse. The lowest values are recorded for the item 15 (*Sees task through to the end, good attention span*) in the hyperactivity/inattention subscale ($\hat{\beta}_{15} = 0.348$), for the item 20 (*Gets on better with adults than other children*) in the Peer relationship problems dimension ($\hat{\beta}_{20} = 0.377$) and for the item 25 (*Often volunteers to help others*), in the Prosocial behaviour subscale with $\hat{\beta}_{25} = 0.537$. These results imply that for both parents and teachers it is ''easy'' to admit that children have good attention span, see chores or homework through to the end, get along better with adults than other children and are often volunteers to help others.

On the other hand, the check of location order displayed in Table 1 reveals that the item ''*Steals from home or school or elsewhere*'', in the Conduct problems subscale, is the most difficult ($\hat{\beta}_{10} = 3.578$), meaning that it is hard to affirm that a son or a pupil steals. Other statements with the highest rating refer to the item 3 (*Often unhappy, depressed or tearful*), ($\hat{\beta}_3 = 2.362$) in the Emotional symptoms subscale and to the item 17 (*Has at least one good friend*) ($\hat{\beta}_{17} = 2.212$), in the Peer relationship problems dimension.

To continue the analysis, we may specify a model allowing for a fixed effect for each subscale of the SDQ. In such a way, we are able to describe and rank the SDQ dimensions according to their different level of ''severity''.

TABLE 1. - *Fixed effects estimates of the items*

Dimension	code	Item Description	$\hat{\beta}_i$	Std. Error	Z-value	Pr(>|z|)	Sig.
Emotional symptoms	1	Often complains of headaches, stomach aches or sickness	1.680	0.128	-13.105	< 2e-16	***
	2	Many worries or often seems worried	1.736	0.130	-13.396	< 2e-16	***
	3	Often unhappy, depressed or tearful	2.362	0.150	-15.749	< 2e-16	***
	4	Nervous or clingy in new situations, easily loses confidence	0.721	0.113	-6.393	1.62E-10	***
	5	Many fears, easily scared	1.075	0.117	-9.210	< 2e-16	***
Conduct problems	6	Often has temper tantrums or hot tempers	1.976	0.136	-14.494	< 2e-16	***
	7	Generally obedient, usually does what adults request	0.783	0.113	-6.912	4.76E-12	***
	8	Often fights with other children or bullies them	1.808	0.132	-13.751	< 2e-16	***
	9	Often lies or cheats	1.546	0.125	-12.354	< 2e-16	***
	10	Steals from home or school or elsewhere	3.578	0.224	-15.973	< 2e-16	***
Hyperactivity/ inattention	11	Restless, overactive, cannot stay still for long	0.976	0.115	-8.452	< 2e-16	***
	12	Constantly fidgeting or squirming	1.132	0.118	-9.627	< 2e-16	***
	13	Easily distracted, concentration wanders	0.618	0.112	-5.524	3.32E-08	***
	14	Thinks things out before acting	0.034	0.110	-0.312	0.754	
	15	Sees task through to the end, good attention span	0.348	0.110	-3.151	0.0016	**
Peer relationship problems	16	Rather solitary, tends to play alone	1.612	0.127	-12.733	< 2e-16	***
	17	Has at least one good friend	2.212	0.144	-15.337	< 2e-16	***
	18	Generally liked by other children	1.680	0.128	-13.105	< 2e-16	***
	19	Picked on or bullied by other children	2.041	0.138	-14.748	< 2e-16	***
	20	Gets on better with adults than other children	0.377	0.110	-3.416	0.0006	***
Prosocial behaviour	21	Considerate of other people's feelings	1.559	0.125	-12.430	< 2e-16	***
	22	Shares readily with other children, for example toys, food	1.155	0.118	-9.793	< 2e-16	***
	23	Helpful if someone is hurt, upset or feeling ill	0.762	0.113	-6.740	1.59E-11	***
	24	Kind to younger children	1.546	0.125	-12.354	< 2e-16	***
	25	Often volunteers to help others (parents, teachers, children)	0.537	0.111	-4.824	1.41E-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Before doing so, we follow a confirmative approach, to evaluate and compare the relative fit of a MIRT model, based on the original five factor structure of SDQ (emotional, conduct, hyperactivity-inattention, peer relationship, Prosocial behaviour) with an another MIRT formulation which considers a lesser number of SDQ subscales.

We examined the underlying components of SDQ through the R mirt package (Chalmers, 2012), freely downloadable from http://cran.r-project.org.

Our attempt is line with previous researches that, by employing a CFA do not offer unequivocal support for the original five-factor model of the SDQ (Dickey and Blumberg, 2004; Goodman *et al*., 2010; Sanne, Torsheim, Heiervang and Stormark, 2009).

A possible alternative to be tested, which finds justification on theoretical grounds, would combine the emotional and peer relationship items into an ''internalizing'' subscale and the conduct and hyperactivity items into an ''externalizing'' subscale (Goodman *et al*., 2010). Our confirmative analysis suggests that the simplification provided by replacing the emotional, peer, conduct, and hyperactivity subscales with internalizing and externalizing factors might be appropriate.

Information criteria (log-likelihood, Akaike Information Criterion and Bayesian Information Criterion), summarized in Table 2, were initially computed to motivate this choice.

TABLE 2. - *Information criteria for MIRT models for SDQ items*

Models	LogLik	BIC	AIC	AICc
Model A (Five Factor model)	-6373.343	13060.05	12846.69	12857.40
Model B (Three Factor model)	-6277.718	12868.80	12655.44	12.666.15

The adequacy of the three factor structure is also confirmed by the usual multiple indices of goodness-of-fit (GOF), like the Comparative Fit Index (CFI), Tucker-Lewis Fit Index (TLI) and the Root-Mean- Square Error of Approximation (RMSEA) (Hu and Bentler, 1999; Marsh, Hau and Wen, 2004).

The conventional rule of thumb is that for a model with an acceptable fit is required that RMSEA should be between 0.05 and 0.08 whilst CFI and TLI should exceed 0.90.

GOF statistics indices revealed that the three factor solution fit the data reasonably better than the original five factor structure (three factors model: RMSEA=0.07, CFI=0.90, TLI=0.89; five factors model: RMSEA=0.07, CFI=0.86, TLI=0.88).

For the identified three-factor model ($D = 3$), we present in Table 3 the fixed effects estimates ($\hat{\beta}_d : d = 1, ..., 3$) for each dimension.

Note that all the estimates are found to be statistically significant. The interpretation of the results in Table 3 suggests that scales displaying the higher values are those deemed less problematic by teachers and parents. It follows that, in the context of the present school, the highest estimates are recorded for the Internalizing subscale

($\hat{\beta}_2 = 1.437$). Equally, from the empirical results it comes out that the prosocial behaviour is the area perceived as the most severe ($\hat{\beta}_3 = 1.054$).

TABLE 3. - *Fixed effects estimates for each dimension*

Dimension	$\hat{\beta}_d$	**Std. Error**	**Z-value**	**Pr(>|z|)**	**Sig.**
Externalizing	1.100	0.058	-18.98	<2e-16	***
Internalizing	1.437	0.059	-24.20	<2e-16	***
Prosocial behaviour	1.054	0.066	-15.89	<2e-16	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

5.1 *Results for the models with temporal trend*

Given the previously mentioned attractive general features of the LLRA models, we continue specifying a model which includes the estimates of the temporal trend, and differentiating between parents and teachers scores. The estimation of the LLRA models has been performed through the eRm R package (Mair and Hatzinger, 2008). In such a way, we can assess for each dimension the measurement changes over time and also the differences between teachers and parents evaluations. For a LLRA model, it is necessary to have the same statistical units assessed in the initial and follow up occasion. Accordingly, we used 117 questionnaires from the parents and 142 from the teachers.

Recall that the Fisher's LLRA procedure, formally grounded in IRT, departs fundamentally from the unidimensionality assumption (Fischer, 1995) and involves the estimation of D change parameters, i.e. the effects of changes occurring across two different measurement occasions ($\tau_d : d = 1, 2, ..., D$).

Following the structure of the previous section, we estimated (for both parents and teachers) two models, one with the original five dimensions ($D = 5$) of the SDQ (emotional, conduct, hyperactivity-inattention, peer relationship, prosocial behaviour) and one with three dimensions ($D = 3$), labelled as said before ''internalizing'', ''externalizing'' and ''prosocial''.

A likelihood ratio test between the two models reveals that regarding the trend parameters there are not differences in the two models (Model for parents' group: Likelihood ratio statistic 3.565, df = 2, p = 0.168. Model for teachers' group: Likelihood ratio statistic 4.027, df = 2, p = 0.134), and the results of the parameter estimates which we discuss below confirm these findings.

Looking at results for the five dimensions model ($D = 5$) for the parents group, the empirical findings suggest that the statistical evaluation of the change is significant only for the hyperactivity/inattention SDQ subscale (Table 4).

The negative value of the trend parameter estimate ($\hat{\tau}_3 = -0.327$) denotes an improvement in that area, meaning that many parents had chosen the lower response categories rather than the higher ones at the follow-up measurement time compared to

the first measure. Even if not statistically significant, it is worth noting that the emotional symptoms and the prosocial behaviour show an improvement.

TABLE 4. - *Trend parameter estimates for* D=5

		$\hat{\tau}_d$	Std.Err	Lower CI	Upper CI	Sig.
Parents	emotional symptoms	-0.230	0.1701	-0.5630	0.1040	
	conduct problems	0.062	0.1755	-0.2820	0.4060	
	hyperactivity/inattention	-0.327	0.1666	-0.6540	-0.0010	***
	peer relationship problems	0.000	0.1591	-0.3120	0.3120	
	prosocial behaviour	-0.130	0.1700	-0.4630	0.2040	
Teachers	emotional symptoms	0.405	0.1757	0.0610	0.7500	***
	conduct problems	-0.336	0.2213	-0.7700	0.0970	
	hyperactivity/inattention	0.092	0.1518	-0.2050	0.3890	
	peer relationship problems	0.717	0.1905	0.3440	1.0910	***
	prosocial behaviour	0.438	0.1522	0.1390	0.7360	***

Signif. codes: '***' 0.05

A more detailed analysis focuses on the trend of each item difficult parameter estimate ($\hat{\tau}_i : i = 1, 2, ..., 25$) and reveals that an important contribution was provided by item 15 (*Good attention span, sees chores or homework through to the end*). In particular, the negative value ($\hat{\tau}_{15} = -0.981$) means that there has been an improvement of the parents' opinions, collected on this issue. Also, all the other items of the same dimension show an improvement.

Other notable finding, are detected by examining the trend parameter estimates related to the teachers' group. Three out of five dimensions of SDQ exhibit statistically significant positive estimates implying a worsening passing from the initial to the follow up version of the SDQ. In particular, the empirical evidence is that teachers were more likely to notice peer relationship problems worsening ($\hat{\tau}_4 = 0.717$), followed by changes on prosocial behaviour problems ($\hat{\tau}_5 = 0.438$) and on emotional scale ($\hat{\tau}_1 = 0.405$). Not statistically significant changes were found on the conduct problems and on hyperactivity/inattention subscale (Table 4).

As it has done for the parents' scores, it could be useful here to go into details of the results, looking at the estimates of trend parameters for the single items ($\hat{\tau}_i : i = 1, 2, ..., 25$). Accordingly, for the Peer relationship problems subscale we find out changes between baseline and follow up scores mainly in the measurement of item 16 (*Rather solitary, tends to play alone*) and of the item 20 (*Gets on better with adults than other children*). Respectively, we have $\hat{\tau}_{16} = 1.224$ and $\hat{\tau}_{20} = 1.322$ all with a positive value, revealing a worsening. Finally, also item 17 (*Has at least one good friend*) is statistically significant and with a positive value ($\hat{\tau}_{17} = 0.944$), so contributing to the worsening of the Peer relationship subscale.

In the prosocial scale, the items that most influence the final judgment are those linked to the sharing and to the relationship with other children: item 22 (*Shares readily with other children, for example toys, food*) and item 24 (*Kind to younger*

children). The estimates of their trend parameters are, respectively, $\hat{\tau}_{22} = 0.887$ and $\hat{\tau}_{24} = 0.693$, both in the sense of worsening. Finally, for the emotional scale, the item 2 (*Many worries or often seems worried*) records a positive estimate, $\hat{\tau}_2 = 1.299$, which explains the worsening in this dimension.

Table 5 lists the trend parameter estimates for the models with $D = 3$ dimensions, together with standard errors and 95% confidence intervals. To derive the significance of the parameter estimates it is worth noting that the hypothesis that the parameter coefficient is equal to 0 at the (1-α)100% level of significance is rejected if the (1-α)100% confidence interval does not include 0.

TABLE 5. - *Trend parameter estimates for* D=3

		$\hat{\tau}_d$	**Std.Err**	**Lower CI**	**Upper CI**	**Sig.**
Parents	Internal	-0.107	0.116	-0.3350	0.1200	
	External	-0.144	0.120	-0.3800	0.0920	
	Prosocial behaviour	-0.130	0.170	-0.4630	0.2040	
Teachers	Internal	0.552	0.129	0.3000	0.8040	***
	External	-0.047	0.125	-0.2910	0.1980	
	Prosocial behaviour	0.438	0.152	0.1390	0.7360	***

Signif. codes: '***' 0.05

For the parents group none of the three parameters are significant, although estimates are all negative, suggesting the tendency by parents to express an improvement, in line with the results of the model with five dimensions.

From the teachers side, the trend parameter of the internal subscale ($\hat{\tau}_1 = 0.552$) and that one of the prosocial behaviour subscale ($\hat{\tau}_1 = 0.438$) are significant and positive, meaning a worsening in these two areas. These results are perfectly in line with the model with five dimensions, for which the significant parameters were precisely those of the emotional scale, the peer relationship scales (here included in the internal subscale) and that ones of the prosocial behaviour scale (see Table 4), all with the same sign, meaning a worsening in these areas.

The previous analysis of the two models (with 5 and 3 dimensions) confirm that there is no difference in terms of temporal trend. In both cases, the main result is that while the teachers tend to detect a worsening in some aspects of the SDQ, the parents tend to express an improvement between the beginning and the end of the school year.

6. DISCUSSION AND CONCLUSIONS

This article presented an IRT approach to analysing item response data arising from the SDQ questionnaire. Appropriate interpretations of SDQ scores are essential in providing preventive treatment of child's behavioural and emotional problems which may create substantial distress for the child and his family. For instance, researches

in this field document that behavioural problems in childhood may lead to early substance abuse, depressive symptoms, as well as adolescent delinquency (Heiervang, Stormark, Lundervold, Heimann, Goodman, Posserud, Ullebo, Plessen, Bjelland, Lie and Gillberg, 2007).

There is a large body of literature available on IRT research in general, but only a few studies have directly applied the IRT models to measure the outcomes of SDQ.

The current study represents the first attempt to employ generalised Rasch models with multidimensional latent trait parameters to measure both the SDQ scores and the change in the IRT models parameter estimates which occurred in two different temporal administrations of the questionnaire. In addition, for the first time the SDQ has been included in a school project involving the practice of physical activities according to the principles of eastern martial arts.

More specifically, the statistical procedure applied in this paper relies on the parents' and teachers' assessment of emotional and behavioural problems in a sample of children aged 6 to 10 years, who participated in a school project of karate.

It is widely recognized that different informants' ratings may contain more information than a single informant (Goodman, Ford, Corbin and Meltzer, 2004). Therefore, we followed the multi-informant approach to the evaluations of psychological problems in school children which is deemed as the best-practice strategy (Mohd and Waheeda, 2014).

Essentially, the data analysis involved different steps. Initially, by formulating a descriptive IRT model (one parameter logistic model) for the entire data set (initial and follow up item response data, collected combining teachers' and parents' scores) we have evaluated each single item, obtaining a differentiation according to their ''severity''. This result emphasizes one of the most important feature of the IRT analysis, differing from classical test theory which, on the contrary, gives to all the items the same weight.

Then, in agreement with other studies concerning the evaluation of the latent structure of the SDQ, we compared the original five-factor structure with an alternative three subscales formulation.

Our analyses supported by simplification produced by replacing the emotional, peer, behavioural and hyperactivity subscale with internalizing and externalizing factors.

Next, by specifying an IRT model which allows for a fixed effect for each subscale, we also identified a differentiation of the three dimensions, according to their level of ''severity''. This could be useful for planning possible interventions in the most problematic areas. For example, it is emerged that both teachers and parents identify the problems covered by the prosocial behaviour subscale as the most relevant.

Further, noteworthy results are obtained by exploiting the properties of the LLRA models, which, as discussed in Section 4, address the issue of changes that occur in the time elapsed between the initial and follow-up version of the SDQ. This allows to evaluate, at least in part, the extent to which the project has had an effect on the pupils' behaviour.

The findings of the LLRA model have also been split by differentiating the evaluations of the teachers from those of the parents.

This aspect is very useful, because allows the analysis of the SDQ dimensions from two different points of view. Our results confirm that teachers and parents perceive differently the pupils' behavioural changes.

Overall, compared to parents, teachers are more likely to detect modifications in emotional, peer relationship and prosocial problems or, if we want to refer to the model with three dimensions, teachers reveal a worsening in the internal and prosocial behaviour subscales.

These findings could be caused by different factors, principally the fact that children behave differently in diverse settings (home and school) (de Nijs, Ferdinand, De Bruin, Dekker, Van Dujin and Verhulst, 2004).

Diverse informants (in our case teachers and parents) are likely to remember and, therefore, report problems, differently, according to their perspective of a child (De Los Reyes and Kazdin, 2005). On one hand, teachers are good informants about children's school behaviour and performance and they might evaluate the child relative to the other children in the class.

On the other hand, since parents may usually only observe their own child, it follows that their assessment of SDQ dimensions suffers the lack of an easy access to a comparison to their child or a standard for normal behaviour (problem of a common standard). Besides, even if the parents are more informed on any problems with the child's behaviour in different environments (home, with friends etc...) than we would expect from the teachers, they are more prone to emphasize the positive part of their child's behaviour and underreport problematical behaviours (Kristoffersen and Smith, 2013).

Our findings are in line with the above mentioned literature, as confirmed by looking at the trend parameter estimates (see Table 3) for the last two subscales (prosocial behaviour and peer relationship problems).

Overall, the use of IRT models and in particular of LLRA as proposed here was found useful to ensure an accurate measurement of SDQ scores and a reliable assessment of their changes over the initial and follow-up versions of this survey instrument.

However, a weak point in our approach is linked to the lack of a sample of controls that would allow to manage other factors that may influence the outcomes (such as children and/or teachers distress).

REFERENCES

Achenbach T., Becker A., Dopfner M., Heiervang E., Roessner V., Steinhausen H.C. Rothenberger A. (2008). Multicultural assessment of child and adolescent psycho-pathology with ASEBA and SDQ instruments: research findings, applications, and future directions. *Journal of Child Psychologyand Psychiatry,* **49**, 251-275.

Adams R.J., Wilson M., Wang W.C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, **21**, 116-123.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders*, 4th ed. (DSM-IV). American Psychiatric Association, Washington.

Aschieri P. (2005). *Manuale teorico-pratico di Karate: scuola elementare e media di primo e secondo grado*. Scuola Nazionale FIJLKAM nº 18 collana FIJLKAM, Roma.

Banks A.L. Reed J.A. (2003). Applying mass media to self-defense instruction in physical education. *The Journal of Physical Education, Recreation & Dance,* **74**, 41-45.

Bates D., Maechler M. and Bolker B. (2013). lme4: Linear mixed-effects models using S4 classes. URL http://CRAN.R-project.org/package=lme4. R package version 0.999999-2.

Becker A., Steinhausen H.C., Baldursson G., Dalsgaard S., Lorenzo M.J., Ralston S.J., Döpfmer M., Rothemberger A. (2006). Psychopathological screening of children with ADHD: strengths and Difficulties Questionnaire in a pan-European study. *European Child & Adolescent Psychiatry*, **15(Suppl. 1)**, 56-62.

Beretvas S.N., Kamata A. (2005). The multilevel measurement model: introduction to the special issue. *Journal of Applied Measurement*, **6**, 247-254.

Brown D., Johnson A. (2000). The social practice of self-defense martial arts: applications for physical education. *Quest*, **52**, 246-259.

Chalmers R.P. (2012). MIRT A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, **48(6)**, 1-29.

de Ayala R.J. (2009). *The theory and practice of item response theory*. The Guilford Press, New York.

De Boeck P., Wilson M. (2004). *Explanatory item response models: a generalized linear and nonlinear approach.* Springer-Verlag, New York.

De Los Reyes A., Kazdin A.E. (2005). Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework and recommentations for future studies. *Psychological Bulletin,* **131**, 483-509.

De Mars C. (2010). *Item response theory. Understanding statistics measurement*. Oxford University Press, Oxford.

de Nijs P.F.A., Ferdinand R.F., De Bruin E.I., Dekker M.C.J., Van Dujin C.M., Verhulst F.C. (2004). Attention-deficit/hyperactivity disorder (ADHD): Parents' judgement about school, teachers' judgement about home. *European Child & Adolescent Psychiatry*, **13**, 315-320.

Dickey W.C., Blumberg S.J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of American Academy of Child and Adolescence Psychiatry,* **43**, 1159-1167.

Di Riso D., Salcuni S., Chessa D., Raudino A., Lis A., Alto G. (2010). The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children. *Personality and Individual Differences*, **49**, 570-575.

Di Zio S. (2009). Future strategy for reducing violence against women: the Italian experience. *Foresight*, **12**, 80-91.

Embretson S.E., Reise S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ. Erlbaum.

Fabbri E., Fabbri S., Primi F. (2001). *Educazione psicomotoria e strutturazione dello schema corporeo*. Società Stampa Sportiva, Roma.

Fischer G.H. (1974). *Einführung in die theorie psychologischer tests: grundlagen und anwendungen.* H. Huber.

Fischer G.H., Ponocny I. (1995). Extended rating scale and partial credit models for assessing change. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (pp. 353-370). Springer-Verlag, New York.

Fischer G.H. (1995). Linear logistic models for change. In G.H. Fischer and I.W. Molenaar (Eds.), *Rasch models, recent developments and applications.* Springer-Verlag, New York.

Fox J.P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software,* **20**, 1-16.

Gardner H. (1983). *Frames of mind: the theory of multiple intelligences*. Basic Books, New York.

Giannakopoulos G., Tzavara C., Dimitrakaki C., Kolaitis G., Rotsika V., Tountas Y. (2009). The factor structure of the Strengths and Difficulties Questionnaire (SDQ) in Greek adolescents. *Annals of General Psychiatry*, **8**, 20.

Goodman R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology and Psychiatry*, **38**, 581-586.

Goodman R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry*, **40**, 1337-1345.

Goodman R. Meltzer H., Bailey V. (1998). The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. *European Child and Adolescent Psychiatry*, **7**, 125-130.

Goodman R., Ford T., Simmons H., Gatward R., Meltzer H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, **177**, 534-539.

Goodman R., Ford T., Corbin T., Meltzer H. (2004). Using the Strengths and Difficulties Questionnaire (SDQ) multi-informant algorithm to screen looked after children for psychiatric disorders. *European Child and Adolescent Psychiatry,* **2**, 25-31.

Goodman A., Lamping D.L., Ploubidis G.B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): data from British parents, teachers and children. *Journal of Abnormal Child Psychology*, **38**, 1179-1191.

Hagquist C. (2007). The psychometric properties of the self-reporting SDQ: an analysis of Swedish data based on the Rasch model. *Personality and Individual Differences*, **43**, 1289-1301.

Hatzinger R., Rusch T. (2009). IRT models with relaxed assumptions in eRm: a manual-like instruction. *Psychology Science Quarterly*, **51**, 87-120.

Hawes D.J., Dadds M.R. (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian and New Zealand Journal of Psychiatry*, **38**, 644-651.

He J.P., Burstein M., Schmitz A., Merikangas K.R. (2013). The Strengths and Difficulties Questionnaire (SDQ): the factor structure and scale validation in U.S. adolescents. *Journal of Abnormal Child Psychology,* **41(4)**, 583-595.

Heiervang E., Stormark K.M., Lundervold A.J., Heimann M., Goodman, R., Posserud M., Ullebo A.K., Plessen K.J., Bjelland I., Lie S.A., Gillberg C. (2007). Psychiatric disorders in Norwegian 8- to 10-year-olds: an epidemiological survey of prevalence, risk factors, and service use. *Journal of the American Academy of Child and Adolescent Psychiatry*, **46**, 438-447.

Hu L., Bentler P.M. (1999). Cut-off criteria for fit indices in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, **6**, 1-55.

Janssens A., Deboutte D. (2009). Screening for psychopathology in child welfare: the Strengths and Difficulties Questionnaire (SDQ) compared with the Achenbach System of Empirically Based Assessment (ASEBA). *European Child and Adolescent Psychiatry*, **18**, 691-700.

Kamata A. (2001). Item analysis by the hierarchical generalized linear model. *Journal Educational. Measurement,* **38**, 79-93.

Kristoffersen J.H.G., Smith N. (2013). Gender differences in the effects of behavioural problems on school outcomes. *Discussion Paper Series IZA,* N.7410.

Lakes K.D., Hoyt W.Y. (2004). Promoting self-regulation through school-based martial arts training. *Journal of Applied Developmental Psychology*, **25**, 283-302.

Maier K.S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics*, **26**, 307-330.

Mair P., Hatzinger R. (2008). Erm: Extended Rasch Modeling. *R-package version* 0.10-1.

Marsh H.W., Hau K.T., Wen Z.L. (2004). In search of golden rules: comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu and Bentler (1999) findings. *Structural Equation Modeling*, **11**, 320-341.

Marzocchi G.M., Capron C., Di Pietro M., Duran Tauleria E., Duyme M., Frigerio A., Gaspar M.F., Hamilton H., Pithon G., Simoes A., Therond C. (2004). The use of the Strengths and Difficulties Questionnaire (SDQ) in Southern European countries. *European Child and Adolescent Psychiatry,* **13**, 40-6.

Mohd A.P., Waheeda K., (2014). Multi-informant reporting of behavioural and emotional problems of school students. *Delhi Psychiatry Journal,* **17**, 100-106.

Nosanchuk T.A. (1981). The way of the warrior: the effects of traditional martial arts training on aggressiveness. *Human Relations*, **34**, 435-444.

Reckase M.D. (2009). *Multidimensional item response theory*. Springer, New York.

Reilly E., Friesen R. (2001). Incorporating self-defence into the physical education. *Strategies*, **14**, 14-17.

Saibene F., Rossi B., Cortili G. (1995). *Fisiologia e psicologia degli sport*. Mondadori, Milano.

Sann B., Torsheim T., Heiervang E., Stormark K.M. (2009). The Strengths and Difficulties Questionnaire in the Bergen child study: a conceptually and methodically motivated structural analysis. *Psychological Assessment*, **21**, 352-364.

Smedje H., Broman J.E., Hetta J., von Knorring A.L. (1999). Psychometric properties of a Swedish version of the ''Strengths and Difficulties Questionnaire''. *European Child & Adolescent Psychiatry*, **8(2)**, 63-70.

Stokes M., Mellor D., Yeow J., Hapidzal N. (2014). Do parents, teachers and children use the SDQ in a similar fashion? *Quality & Quantity*, **48**, 983-1000.

Theeboom M., De Knop P. (1999). Eastern martial arts and approaches of instruction in physical education. *European Journal of Physical Education*, **4**, 146-161.

Ullebo A.K., Posserud M.B., Heiervang E., Gillberg C., Obel C. (2011). Screening for the attention deficit hyperactivity disorder phenotype using the Strength and Difficulties Questionnaire. *European Child and Adolescent Psychiatry*, **20**, 451-458.

Vockell E.L., Kwak H.S. (1990). Martial arts in the classroom. *Clearing House*, **64**, 61-64.

Walsh D. (2000). Should martial arts be taught in physical education classes? *The Journal of Physical Education, Recreation, & Dance,* **71**, 12-12.

Woerner W., Fleitlich-Bilyk B., Martinussen R., Fletcher J., Cucchiaro G., Dalgalarrondo P., Lui M., Tannock R. (2004). The Strengths and Difficulties Questionnaire overseas: evaluations and applications of the SDQ beyond Europe. *European Child and Adolescent Psychiatry,* **2**, 47-54.