Sport creates Change

Impacts of Sport for Development on Gender Equality, Psychosocial Wellbeing and Social Cohesion in the Kurdistan Region of Iraq

giz Deutsche Gesellschaft
für Internationale
Zusammenarbeit (GIZ) GmbH

IMPRINT

Content

List of Figures and Tables

List of Abbreviations

BMZ   German Federal Ministry for Economic Cooperation and Development
FWE   Friends of Waldorf Education
GIZ   Gesellschaft für Internationale Zusammenarbeit
IDP   Internally Displaced Person
ISIS   Islamic State of Iraq and Syria
KRI   Kurdistan Region of Iraq
MHPSS   Mental Health and Psychosocial Support
NGO   Non-Governmental Organisation
PE   Physical Education
PTSD   Post-Traumatic Stress Disorder
RCT   Randomized Controlled Trial
SDG   Sustainable Development Goal
S4D   Sport for Development
VAWG   Violence Against Women and Girls
WHO   World Health Organisation

ABSTRACT

The quasi-experimental impact study in the Kurdistan Region of Iraq was carried out as a collaboration between two programs of the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ): the global program "Sport for Development" and the regional project "Exchange, Education and Conflict Management through Sport for Development Jordan/Iraq", and the German Sport University Cologne. The aim was to examine the contribution of Sport for Development on specific targets of the Sustainable Development Goals (SDGs), namely SDG Target 3.4, SDG Target 5.1, 5.2, 5.4 and SDG Target 10.2. The target group consists of children and youth between the ages of 10 and 19 in camps for internally displaced persons (IDPs) and host communities in the Kurdistan Region of Iraq.

Results show that Sport for Development contributes to the achievement of SDG Target 3.4: Sport for Development measures have a statistically significant, medium effect on psychosocial wellbeing of children and youth and thus increase psychosocial wellbeing. Sport for Development also shows a statistically significant, albeit weak effect on social-emotional skills. Mixed sport teams increase the psychosocial wellbeing of boys more compared to all-boys teams. The study also shows that psychosocial wellbeing of children and youth in IDP camps is lower than that of children and youth from host communities. However, Sport for Development increases the wellbeing of the former more than the latter.

Sport for Development has a statistically significant, strong effect on SDG Target 5.1 and improves attitudes towards gender equality in the studied context. Sport for Development has a statistically significant, medium to strong effect on SDG Target 5.2 and reduces the acceptance of and propensity for gender-based violence against women and girls. The analyses also show a statistically significant, medium effect on SDG Target 5.4: Sport for Development measures promote attitudes towards an equal division of labour in the household – but do not contribute to behavioural changes in this regard throughout the household. The breaking down of social norms towards an equal division of labour requires a holistic approach, which Sport for Development alone cannot achieve.

The contribution of Sport for Development to SDG Target 10.2 was also examined. A statistically significant, medium effect could be demonstrated: Sport for Development strengthens inclusive behaviour towards other religions and ethnicities and promotes social cohesion.

After the defeat of the so-called Islamic State of Iraq and Syria (ISIS) in late 2017, the number of internally displaced persons returning to their area of origin remains relatively low, due to ongoing conflicts in Sinjar, the lack of employment and prospects for livelihoods. Hence, as of November 2022, over 1.2 million people are internally displaced within their own country, thereof 700,000 residing in the Kurdistan Region of Iraq (KRI). Nearly 180,000 persons are living in 26 camps for internally displaced persons (IDPs) across the governorates of the Kurdistan Region of Iraq and the Ninewa Governorate. 15 IDP camps are located in Dohuk Governorate (REACH, 2021; UNHCR, 2022).

1.2 million internally displaced persons in Iraq, with 700,000 living in the Kurdistan region of Iraq.

In the Kurdistan Region of Iraq, the situation is comparatively stable, yet characterized by high levels of poverty, unemployment, and limited access to services. These conditions led to an increase in mental health problems, including depression, anxiety, and post-traumatic stress disorder (PTSD). Additionally, the protracted crisis and instability in the region, aggravated by the COVID-19 pandemic, created significant stressors that had adverse effects on the mental health of IDPs (ISTOR, n.d.).

The mental health challenges faced by IDPs in the Kurdistan Region of Iraq are exacerbated by the lack of mental health services in the region. According to the World Health Organisation (WHO) there are only 0.3 psychiatrists per 100,000 people in Iraq, compared to the global average of 1.2 psychiatrists per 100,000 people (WHO, 2019). Additionally, there is a significant shortage of trained mental health professionals, with many working in other fields due to the lack of mental health funding and resources. Despite these challenges, there are several initiatives underway to improve the mental health situation of IDPs. The WHO and other international organisations have launched mental health programs that provide counselling and psychosocial support to IDPs. These programs aim to increase access to mental health services and reduce the stigma associated with mental illness.

War and displacement left IDPs but also the Kurdish host communities with mistrust and prejudices. A current literature review conducted by the German Institute of Development and Sustainability shows that conflict harms social cohesion (Fiedler & Rohles, 2021). Using the definition by Leininger et al. (2020), social cohesion refers to the vertical and horizontal relations among members of society and the state – with three core dimensions: trust, cooperation for the common good, and an inclusive identity. In the present study, the horizontal relations are of interest.

One important characteristic of social cohesion is trust between members of society (horizontal/ social trust). It means the "ability to trust people outside one's familiar or kinship circles" (Mattes & Moreno, 2018, p. 357) and can also be referred to as "bridging trust" that acts as the "bond that people share across a society and across economic and ethnic groups, religions, and races" (Rothstein & Uslaner, 2005, p. 45). A large number of studies suggests that conflict and war decrease social trust (Fiedler & Rohles, 2021). This also illustrates the interlinkages

between psychosocial wellbeing and social cohesion: If PTSD symptoms persist after war, victims are expected to reduce interpersonal exchange and relationships with others, as well as withdraw from social activities, resulting in a reduction in both cooperation and trust. More recent research suggests that war and conflict decrease cooperation, especially out-group cooperation, which negatively affects social cohesion as a whole. Looking at the third dimension, a socially cohesive society is characterized by peaceful co-existence of individuals with different identities without the domination of one particular identity over the collective identity. Existing research shows that conflict and war increase group identification but decrease national identity (Fiedler & Rohles, 2021).

Conflict and war are not only affecting social cohesion but also some of the most vulnerable groups: Women and girls. War and displacement negatively affect the security situation and safety of women and girls (UN WOMEN, MADRE, City University of New York, 2021). In Iraq, gender-based discrimination against women and girls is heavily impeding their opportunities for social and political participation as well as their access to education, health care and employment. This discrimination is based on the patriarchal structure of society, gender stereotypes and unequal power relations. Gender-based violence against women and girls (VAWG) is thereby the most extreme expression of gender inequality.

While the Iraqi constitution forbids any form of violence within families, it also considers the punishment of a wife by her husband to be a legal right. In the KRI, a specific family law was passed in 2011 granting women statutory protection against gender-based violence declaring it a criminal act. However, there is insufficient political motivation to enforce this right to protection. In central Iraq and the KRI, any approved religious group can settle issues relating to family and civil status. The often conservative interpretation of laws fosters gender-based violence and discrimination against women and girls. Especially vulnerable groups such as female refugees and internally displaced women and girls are severely affected by gender-based violence. Additionally, the Covid-19 pandemic saw an increase of violence within families and gender-based VAWG: Two thirds of Iraq's social authorities reported an increase of women seeking help. According to a report by the Supreme Judicial Council of Iraq, domestic gender-based VAWG peaked in 2021 with 18.602 lawsuits filed (Supreme Judicial Council of Iraq, 2022). In the first six months of 2022, more than 10.000 cases of domestic violence were registered, almost 8.000 being cases of gender-based VAWG. The rise in gender-based VAWG highlights the limited legal and financial support available to victims and survivors who are stigmatized and often trapped in abusive households due to persisting economic dependencies and conservative social norms that consider it shameful for women to leave the household or to take legal action.

Sport provides a platform to question traditional gender roles and to change these in the long term. Team sport promotes social skills like fairness, tolerance, and mutual understanding through communication, and it helps players overcome cultural gender stereotypes. Players are also given the opportunity to develop healthy, respectful relationships amongst different genders. Sport, when implemented in a pedagogically and socially appropriate manner, can provide a safe and protected environment in which sensitive issues can be discussed freely and without prejudice – with both girls and boys. The empowerment of girls and women, strengthening their self-worth and self-confidence is an important prerequisite for achieving gender equality.

Sport can potentially be used in a way to improve a child's psychosocial wellbeing. Sport for Development (S4D) practices are designed, accompanied, and managed in a trauma-informed manner and supervised by a coach who has been trained and is experienced in providing psychosocial support. If sport activities are supervised properly, they can empower children and (re)build coping strategies. It can give them a chance to experience the following: empathy, fairness, cooperation, self-control and self-efficacy, a sense of belonging to a group, reflection on their emotions and behaviour, coping with both – success and frustration and more.

Sport can play a role in supporting various elements of social cohesion. However, sport is a social activity that might divide as much as it unites. Yet, when implemented in a sensitive and culturally adapted manner, S4D can contribute by building core competences on an individual level while fostering social inclusion on a community level. S4D activities provide an arena for the community to witness cooperation and support building relationships – connecting individuals to communities and across social groups. Essential competences such as critical thinking, respect, tolerance, cooperation and problem-solving can develop. Furthermore, S4D activities require interaction and can thus help challenge stereotypes between different ethnicities or religions and break barriers of mistrust between young people.

S4D coaching focuses on the five principles of:
1. Multidimensional development of participants taking cognitive, social, physical and sport-related development into consideration
2. Developing life skills in a way that allows participants to incorporate and transfer them to their daily life contexts
3. The roles and responsibilities of the coach specifically focusing on the role model function of the coach
4. Appropriate educational goals: activities chosen should neither be too easy nor too difficult for participants, but adjusted to the personal and sociocultural context and background of children and youth in an attempt to manage diversity
5. A clear training structure that is based on coaching high quality sport sessions while integrating social-emotional and life skills in every part of the session (sensitization, warm-up, main-part, conclusion, reflection)
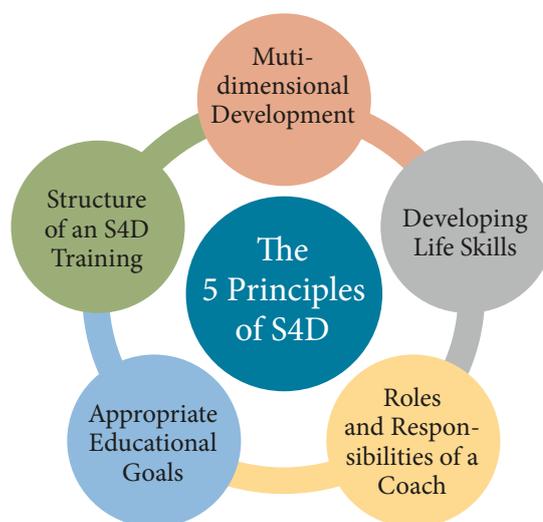
Figure 1:
The 5 Principles of S4D

SPORT FOR DEVELOPMENT IMPLEMENTATION IN THE KURDISTAN REGION OF IRAQ

On behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ), the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH supports the introduction of gender-sensitive sport activities into camps for IDPs as well as host communities in the Kurdistan Region of Iraq to foster psychosocial well-being and a peaceful coexistence. The project started its work in 2018, training local coaches to identify and leverage the potential of sport for children's development. Until now, over 30,000 children and youth have been reached through 320 trained sport coaches in the KRI.

> **320 trained S4D coaches and 30,000 children reached through S4D in the Kurdistan region of Iraq.**

To implement activities on the ground, GIZ partners with the non-governmental organisation (NGO) Friends of Waldorf Education (FWE) that focuses on providing psychological first aid and trauma pedagogy to survivors of ISIS. Over a period of five years, local staff was trained by international emergency pedagogy and trauma pedagogy experts. At the time of the S4D intervention each FWE trauma pedagogy practitioners had completed over 23,000 hours of designing and conducting group based mental health and psychosocial support (MHPSS) activities with children between 3 - 12 years and provided psychosocial advisory services to parents. In order to minimize the impacts of trauma and to stabilize children living in IDP and refugee camps, humanitarian organisations, including FWE, have been attempting to smoothen the gap between need and provision by establishing and running "Child Friendly Spaces" since 2015. These "Child Friendly Spaces" are safe areas dedicated to children and youth who survived potentially traumatic events and are living in hardship conditions. Here, FEW trauma pedagogy practitioners created an environment characterized by structural and emotional safety and continuity.

GIZ started cooperating with FWE in 2019. In 2021/2022, 12 trauma pedagogy practitioners were teamed up with seven physical education (PE) graduates from the University of Zakho. These 19 coaches from different backgrounds are also representing this study's target group: Eight coaches are IDPs, eleven coaches come from the host community. The 19 coaches formed coaching teams based on principles of diversity (gender mixed, IDP & host community background, trauma practitioner & PE graduate) wherever possible. The vast majority of S4D coaches who implemented the S4D activities evaluated in this report are certified S4D coaches since 2019 according to the five principles of S4D above. They have been continuously trained specifically in trauma informed sport coaching, as well as in using sport to foster gender equality and to prevent gender-based VAWG. In total, each of the 19 coaches received more than 110 hours of training.

The S4D implementation being evaluated in this study took place in the three IDP camps Bersive I, Bersive II and Chamishko as well as the neighbouring host community of Zakho in Duhok Governorate. The Camps Bersive I and II are located 20 km east of the district capital Zakho in a rural environment accommodating over 5,000 and 7,000 persons respectively, while the camp Chamishko borders the city of Zakho, hence offering its over 21.000 inhabitants' greater access to infrastructure and facilities. All three camps were established in November 2014, only a few months after the invasion of ISIS in Mosul (June 2014) and Sinjar (August 2014).

20 football teams were formed that started regular trainings sessions twice a week in October/November 2021, either using the small-sided mini pitches at the "Child Friendly Spaces" or bigger tented pitches near the respective camps as well as in the host community of Zakho city. At the same time, 10 of the 19 coaches were selected to receive additional weekly theoretical and practical training sessions on ultimate frisbee as the trend sport is relatively new in Iraq – hence competences had to be established first. Five ultimate frisbee teams (one female team in the host community, four mixed teams in IDP camps) were eventually formed and started training in December 2021. In total, the S4D intervention focusing on football lasted seven months, the intervention using ultimate frisbee lasted six months.

Among the 25 teams, there were 20 teams with IDPs, 5 teams with members of the host community, 11 all-boys teams, 1 all-girls team and 13 mixed teams. Participants were divided into age-appropriate teams: 10-12, 13-15, 16-19 years old. All teams underwent one-hour sessions twice per week. Sessions were specifically designed to foster psychosocial wellbeing, to increase gender equality, and to contribute to preventing gender-based VAWG. Sessions were based on the manual "Sports Builds Gender Equality" (GIZ, 2022) specifically designed for that purpose and focused on contributing positively to identified challenges of children and youth through S4D's social and life skill-based approach. The coaches followed a predetermined curriculum and were supervised by a S4D instructor who conducted multiple field visits. Additionally, coaches kept files on their sessions for the instructor to monitor.

Various events and leagues accompanied the regular trainings sessions:
- A league for all mixed football teams from February – June 2022
- An ultimate frisbee league for all frisbee teams (March – June 2022)
- A "night-league" for all-male football teams during Ramadan 2022 (April 2022)

All leagues followed the principles of "Fair-Play" which is based on an additional scoring system for fairness and team spirit. The "International Day of Sport for Development and Peace" was celebrated on the 6th of April with a tournament of all mixed football teams and the female team, while the ultimate frisbee teams joint a regional frisbee event in Dohuk in May 2022. The all-male football league celebrated the finals in June 2022 with a game festival.

RESEARCH OBJECTIVE

The aim of this study was to evaluate whether GIZ's S4D approach is a meaningful tool to promote gender equality and to prevent gender-based VAWG, to strengthen psychosocial wellbeing and to foster social cohesion. For this purpose, GIZ joint forces with the German Sport University Cologne, the University of Dohuk and the NGO Friends of Waldorf Education. More specifically, the present study evaluated if a six-to-seven-month long S4D intervention in the Kurdistan Region of Iraq contributes to the realization of:

SDG Target 3.4: By 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being.

SDG Target 5.1: End all forms of discrimination against all women and girls everywhere.

SDG Target 5.2: Eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation.

SDG Target 5.4: Recognize and value unpaid care and domestic work through the provision of public services, infrastructure and social protection policies and the promotion of shared responsibility within the household and the family as nationally appropriate.

SDG Target 10.2: By 2030, empower and promote the social, economic and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion or economic or other status.

RESEARCH DESIGN

A quasi-experimental, longitudinal study design was used to examine possible impacts of Sport for Development on children and youth' psychosocial wellbeing (SDG Target 3.4), attitudes and behavioural changes towards gender norms and gender-based violence against women and girls (SDG Target 5.1, 5.2, 5.4) as well as inclusive behaviour (SDG Target 10.2) in the KRI. The intervention group consisted of all 25 football and ultimate frisbee teams who participated in S4D activities twice per week for a duration of six to seven months. The comparison group did not receive any type of treatment and thus was not involved in any type of S4D activity. Data collection took place at two points: the baseline was conducted in November/ December 2021 and the endline in May/ June 2022.

The allocation ratio between intervention and comparison group is 1:1. 317 children and youth were assigned to the intervention group, 319 to the comparison group. In total, 636 children and youth participated in the study. The gender ratio is 28% females in the intervention group and 35% in the comparison group. Initially, the aim was a 50% gender quota but had to be reduced to approximately 30%. The somewhat lower representation of female participants compared to males is due to difficulties in reaching girls for sport projects rooted in cultural and societal norms. GIZ and FWE encountered a lot of rejection and barriers when trying to engage girls for the intervention.

The ratio of IDPs and members of the host community is 4:1 as the focus of the S4D intervention was on IDPs. The project location was predetermined by the GIZ project on site and the local partner due to the already existing "Child Friendly Spaces" in all three camps, access, and the well-established relations to the camp management.

This is a full survey, meaning all children and youth participating in the S4D intervention also participated in the study. To participate, the following criteria were applied for participants of the intervention and comparison group: age 10 to 19, not having participated in any S4D activity or/and FWE trauma pedagogy activity before, being willing and available to participate in S4D activities twice per week (for intervention group) and being from either an IDP camp (Bersive I, Bersive II and Chamishko) or the host community of Zakho.

Since this is a full survey, there was no randomization of a sample. However, the selection of participants in the intervention was done using the non-probability sampling techniques self-selection sampling and non-proportional quota sampling along the above-mentioned criteria. Self-selection sampling was carried out through 19 S4D coaches who disseminated the information about the S4D intervention within the IDP camps. In the host community, coaches went into schools to introduce the S4D project. Thus, all children and youth in the three IDP camps fitting the selection criteria and voluntarily agreeing to be part of the project were registered to participate. In the host community, self-selection sampling was paired with non-proportional quota sampling to achieve a 20% quota of host community participation. One additional criterion for participation for the host community was school classes in the morning to ensure availability in the afternoon for S4D activities. Non-proportional quota sampling was also applied for the gender representation in both comparison and intervention group. There were several dropouts in the first weeks which were replaced with mainly girls in a further attempt to elevate the female participation rate. For the comparison group, children and youth fitting the defined criteria were selected based on self-selection sampling and non-proportional quota sampling.

For the study, a standardized questionnaire was developed by researchers from German Sport University Cologne and GIZ taking into consideration the specific project intervention and local context. Psychologists and gender experts were consulted specifically for the items regarding SDG 3 and SDG 5. After a first pilot, the questionnaire was readjusted to fit the age group and cultural context. A second pilot was conducted which resulted in further adjustments of the questionnaire. The questionnaire was translated into Arabic and Kurdish and digitized. All children and youth of the intervention and comparison group were interviewed by trained interviewers (psychology and PE graduates) via tablets and an offline survey application. Before the endline was conducted, six focus group discussions with the intervention group were held with six participants each (36 in total), mainly focusing on SDG Target 5.1 and 5.2.

In total, 636 children and youth were interviewed. During the baseline, seven participants from the intervention group did not participate in the baseline interviews, but did participate in the endline, reducing the number of the baseline interviewees of the intervention group to 310 instead of 317. The following table illustrates the participant flow between baseline and endline:

	INTERVENTION GROUP	COMPARISON GROUP
TOTAL	317	319
BASELINE	310	319
ENDLINE	236	253

Table 1: Intervention and Comparison Group

The high fluctuation between baseline and endline is due to the unstable environment in IDP camps and due to the fact that the endline was close to the end of the school year with parents taking their children out of school early to work on fields.

The next table illustrates the distribution of the intervention and comparison group in subgroups:

	GENDER		AGE GROUP*			BACKGROUND**	
	FEMALE	MALE	10–12 YEARS	13–15 YEARS	16–19 YEARS	IDP	HOST COMMUNITY
INTERVENTION GROUP	88	229	89	162	56	253	64
COMPARISON GROUP	113	206	101	152	48	240	60

Table 2: Subgroups 1

S4D INTERVENTION GROUP			
TYPE OF SPORT		TYPE OF SPORT TEAM	
FOOTBAL	ULTIMATE FRISBEE	MIXED TEAM	SAME SEX TEAM
253	64	220	97

Table 3: Subgroups 2

*10 missing cases in the intervention group. 18 missing cases in comparison group

**19 missing cases in comparison group

ETHICAL CONSIDERATIONS

The study was approved by GIZ's data protection unit. The anonymity of the participants is guaranteed by GIZ, and the General Data Protection Regulation by the European Union is applied. As the study's target group is minors, approval for participation was sought by parents/ legal guardians beforehand.

While we acknowledge that especially self-selection sampling and using coaches to recruit/select participants contribute to some kind of selection bias, the coaches function as important gate keepers to the target group and without them, the target group would remain inaccessible. Further, probability sampling would have provoked ethical concerns. Making an offer for participation in S4D activities and then randomly selecting participants would at the same time exclude others willing and hoping to participate. Especially in vulnerable contexts such as IDP camps, such practices would create not only large disappointments but also possible conflict on already mentally strained children and youth. While a randomization at cluster level (camps) might be possible in theory, in practice the number and access to such camps remains limited.

RESULTS

The following section outlines the results from the base- and endline survey. Qualitative results are, whenever possible, triangulated with the quantitative results. This is the case for results on attitudes towards gender equality and gender-based VAWG.

Impacts on SDG Target 3.4: Psychosocial Wellbeing

Psychosocial wellbeing being a latent variable, 22 Likert scale items measuring psychological and social wellbeing were combined into one scale. Items were derived from KIDSSCREEN27 and KIDSSCREEN52 and adapted to the local context of IDPs (Ravens-Sieberer et al., 2014). As the items on psychosocial wellbeing have different scales (four to five scales), Cronbach's Alpha on Standardized Items was used to measure the internal consistency of the scale. Cronbach's Alpha on Standardized Items in the baseline is 0.800 and in the endline 0.820 thus showing a very good internal consistency of the scale.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,786	,800	22

Figure 2: Cronbach's Alpha, Psychosocial Wellbeing, Baseline

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,809	,820	22

Figure 3: Cronbach's Alpha, Psychosocial Wellbeing, Endline

To compare the psychosocial wellbeing between the intervention and comparison group and within each group over time, a mixed between-within ANOVA (also called split-plot ANOVA, between-within ANOVA, or mixed factorial ANOVA) was conducted. It was decided not to conduct a MANOVA (Multivariate Analysis of Variance) as the six dependent variables do not necessarily relate to each other: psychosocial wellbeing; social-emotional competences; gender equality (misogynistic attitudes); gender-based violence against women and girls; recognition of unpaid care and domestic work; social cohesion (inclusive behaviour). To reduce the risk of a Type 1 error, a more stringent alpha value was set, and a Bonferroni adjustment used. With six dependent variables, the new alpha value is $p = 0.05 / 6 = 0.008$ to control for Type 1 error across multiple tests.

As SPSS automatically calculates a list-wise case exclusion, meaning if a single value is missing from a variable, the entire case will be excluded from the analysis, 229 cases in the intervention group and 235 cases in the comparison group were taken into consideration by SPSS.

Between-Subjects Factors

		Value Label	N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	229
	2,00	Control group (no S4D)	235

Figure 4: Between-Subjects Factors, Psychosocial Wellbeing

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
PsychosocialWellbeing.1	Intervention group (S4D)	4,2099	,43890	229
	Control group (no S4D)	4,2804	,40946	235
	Total	4,2456	,42525	464
PsychosocialWellbeing.2	Intervention group (S4D)	4,5395	,34990	229
	Control group (no S4D)	4,3236	,43829	235
	Total	4,4301	,41117	464

Figure 5: Descriptive Statistics, Psychosocial Wellbeing

We fulfil the prerequisite for conducting mixed between-within ANOVAs with the dependent variables being interval-scaled and the independent variable/ between-subjects factor nominal-scaled with two independent groups. The within-subjects factor is time (two measuring points) and is independent and nominal-scaled. Regarding outliers, the questions were programmed as Likert scales with pre-defined answer options in order to prevent outliers. As ANOVA is a quite robust analytical method against violations of the normality assumption, especially with large sample sizes and balanced designs, the normality assumption can be neglected (Tabachnik & Fidell 2007; Salkind 2010): The sample size for Psychosocial Wellbeing is equally distributed among both groups and rather large (see Figure 4).

The same applies to variance homogeneity which is tested through Levene's Test of Equality of Error Variances. While variance homogeneity can be neglected with large sample sizes and balanced designs, in our case we still fulfil the assumption of homogeneity. Homogeneity of variances was asserted using Levene's Test based on median which shows that equal variances can be assumed ($p = 0.192$ in the baseline and $p = 0.051$ in the endline). We use Levene's Test based on median as it is more robust then based on mean.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
PsychosocialWellbeing.1	Based on Mean	1,999	1	462	,158
	Based on Median	1,708	1	462	,192
	Based on Median and with adjusted df	1,708	1	461,587	,192
	Based on trimmed mean	1,911	1	462	,168
PsychosocialWellbeing.2	Based on Mean	5,234	1	462	,023
	Based on Median	3,824	1	462	,051
	Based on Median and with adjusted df	3,824	1	427,447	,051
	Based on trimmed mean	4,476	1	462	,035

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + InterControlGroup
   Within Subjects Design: MeasuringPoints

Figure 6: Levene's Test of Equality of Error Variances, Psychosocial Wellbeing

Box's Test of Equality of Covariance Matrices[a]

Box's M	17,584
F	5,834
df1	3
df2	39249857,363
Sig.	<,001

As we have a mixed design, we also check for homogeneity in covariance by using Box's Test of Equality of Covariance Matrices. Since the power of Box's Test is dependent on the number of cases, the test becomes more significant the larger the sample is. Some authors therefore recommend not testing the Box's Test at a .05 significance level, but at .025 or .01 (Mertler, 2004) or .001 (Verma, 2015; Warner, 2012). In our case, the assumption of homogeneity of covariance is violated as $p < 0.001$. However, with a large sample size and a balanced design, we can neglect this assumption.

Figure 7: Box's Test of Equality of Covariance Matrices, Psychosocial Wellbeing

The assumption of sphericity can be neglected, as it only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

The mixed between-within ANOVA conducted to assess the impact of Sport for Development on psychosocial wellbeing across two time periods (pre-intervention, post-intervention) shows a significant interaction between group affiliation and time ($p < 0.001$):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	8,056	1	8,056	66,884	<,001	,126
	Greenhouse-Geisser	8,056	1,000	8,056	66,884	<,001	,126
	Huynh-Feldt	8,056	1,000	8,056	66,884	<,001	,126
	Lower-bound	8,056	1,000	8,056	66,884	<,001	,126
MeasuringPoints * InterControlGroup	Sphericity Assumed	4,757	1	4,757	39,496	<,001	,079
	Greenhouse-Geisser	4,757	1,000	4,757	39,496	<,001	,079
	Huynh-Feldt	4,757	1,000	4,757	39,496	<,001	,079
	Lower-bound	4,757	1,000	4,757	39,496	<,001	,079
Error(MeasuringPoints)	Sphericity Assumed	55,648	462	,120			
	Greenhouse-Geisser	55,648	462,000	,120			
	Huynh-Feldt	55,648	462,000	,120			
	Lower-bound	55,648	462,000	,120			

Figure 8: Tests of Within-Subjects Effects, Psychosocial Wellbeing

Psychosocial wellbeing was evaluated on a scale from one to five. Figure 9 shows a positive increase in psychosocial wellbeing among the intervention group over time and compared to the comparison group. On the contrary, the comparison group shows only a slight increase in psychosocial wellbeing over time. The tendency towards higher values in the response behaviour of children and youth is a well-known phenomenon in social sciences and psychology. On the one hand, this is due to social desirability. On the other hand, children and youth often reflect differently on their response behaviour after an intervention. This behaviour is observed in all our analyses. Looking at the effect size, both Cohen's F and partial eta squared was analysed. Cohen's F = 0.293 and partial eta squared = 0.079 show a medium effect size.

Estimated Marginal Means of MEASURE_1

Which group belongs the interviewee to?
— Intervention group (S4D)
— Control group (no S4D)
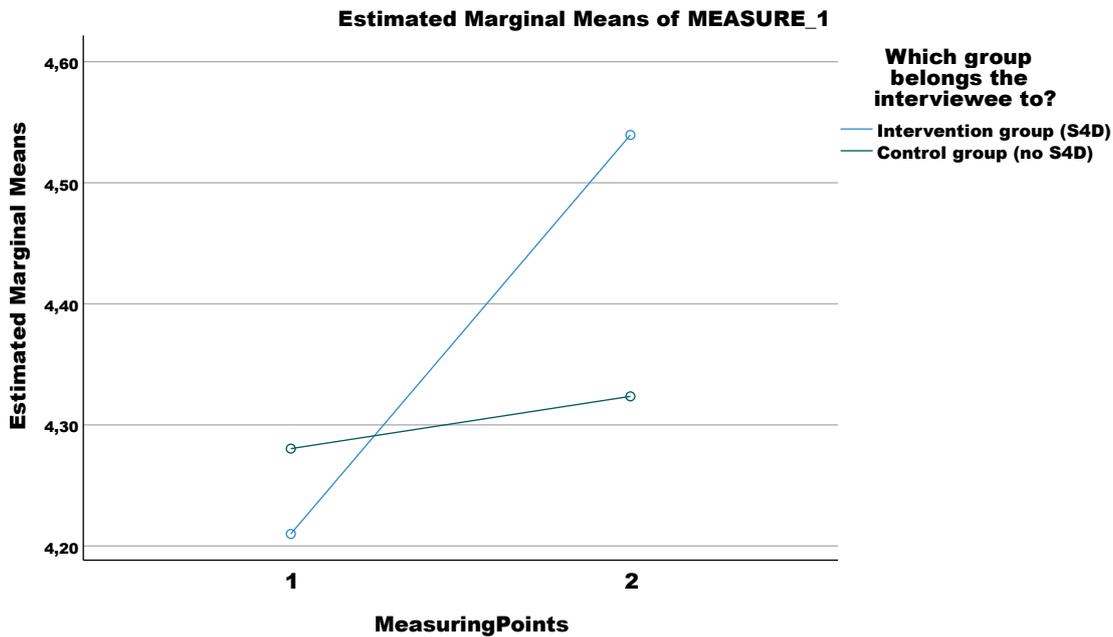
Figure 9: Estimated Marginal Means, Psychosocial Wellbeing

In conclusion, there is a significant interaction between S4D participation and time, with S4D participation increasing psychosocial wellbeing, Greenhouse-Geisser $F(1.00, 462.00) = 39.50$, $p < 0.001$, with a medium effect size (Cohen's $F = 0.293$, partial $\eta^2 = 0.079$).

These results demonstrate that S4D can be used as an effective psychosocial support measure that fosters a sense of belonging and hears children's and youth's voices. In the focus group discussions, this was reflected in pictures drawn by the participants showing how they perceive themselves after the S4D implementation. Pictures revolved around perceived safety, family, friendship, overcoming struggles, inner strength, self-love, trust, and wishes for the future.

The illustrated example here was drawn by a ten-year-old boy who drew a pyramid with each colour representing a period in his life. Darker colours represent dark periods in his life such as the flight from ISIS with his family, their arrival at the IDP camp and not being well. The colours grow lighter with the colour green representing hope and his current feeling. It also represents the S4D pitch where he feels safe and well. The top colour is red and represents love with a heart at the right side. Another participant drew a picture of a caterpillar and a butterfly representing his journey from caterpillar to butterfly, him being halfway there.

Figure 10: Picture from Focus Group Discussion 1

S4D fosters psychosocial wellbeing and has an impact on SDG Target 3.4.

Through S4D, children and youth experience self-efficacy, which is specifically important in the context of forced migration and potential traumatisation where stability, belonging and a healthy environment are often lacking. This makes S4D a meaningful approach not only for sport projects but also for development interventions working on MHPSS – in crisis or post crisis settings.

Psychosocial Wellbeing and its Interlinkage with Gender

It was also analysed within the intervention group if gender, age, type of sport (football or ultimate frisbee), sport team (mixed or same sex team) or background (IDP or host community) have a significant influence on psychosocial wellbeing. While there were no significant impacts of gender, age and type of sport on psychosocial wellbeing found, a significant impact of the background and type of sport team was identified. Since most same sex teams were male teams, it was assumed that mixed teams specifically increase psychosocial wellbeing of male children and youth and that male participants benefit from the presence of female participants regarding atmosphere, social interactions, and group dynamics – all factors contributing to psychosocial wellbeing. This assumption was tested with the sample of male participants and the hypothesis was confirmed:

Between-Subjects Factors

		Value Label	N
typeofgroup.1: Type of group:	1,00	Mixed group	109
	2,00	Single sex group	59

Figure 11: Between-Subjects Factors, Psychosocial Wellbeing, Mixed and Single Sex Groups

Descriptive Statistics

	typeofgroup.1: Type of group:	Mean	Std. Deviation	N
PsychosocialWellbeing.1	Mixed group	4,1070	,39734	109
	Single sex group	4,3251	,42052	59
	Total	4,1836	,41763	168
PsychosocialWellbeing.2	Mixed group	4,6015	,32983	109
	Single sex group	4,4344	,40061	59
	Total	4,5428	,36400	168

Figure 12: Descriptive Statistics, Psychosocial Wellbeing, Mixed and Single Sex Groups

Homogeneity of variances was partly asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.956) but not in the endline (p = 0.006) in the endline). Since variance homogeneity can be neglected with large sample sizes and balanced designs, we decided to neglect this prerequisite.

Levene's Test of Equality of Error Variances[a]

a. Design: Intercept + typeofgroup.1
Within Subjects Design: MeasuringPoints

		Levene Statistic	df1	df2	Sig.
PsychosocialWellbeing.1	Based on Mean	,012	1	166	,914
	Based on Median	,003	1	166	,956
	Based on Median and with adjusted df	,003	1	164,225	,956
	Based on trimmed mean	,004	1	166	,952
PsychosocialWellbeing.2	Based on Mean	10,030	1	166	,002
	Based on Median	7,707	1	166	,006
	Based on Median and with adjusted df	7,707	1	165,869	,006
	Based on trimmed mean	10,135	1	166	,002

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

Figure 13: Levene's Test of Equality of Error Variances, Psychosocial Wellbeing, Mixed and Single Sex Groups

Homogeneity in covariance can be assumed with Box's Test of Equality of Covariance Matrices showing p = 0.282. The assumption of sphericity can be neglected, as it only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

Box's Test of Equality of Covariance Matrices[a]

Box's M	3,874
F	1,272
df1	3
df2	402044,355
Sig.	,282

The mixed between-within ANOVA conducted to assess the impact of different types of sport teams on male participant's psychosocial wellbeing across two time periods (pre-intervention, post-intervention) shows a significant interaction between sport teams and time (p < 0.001):

Figure 14: Box's Test of Equality of Covariance Matrices, Psychosocial Wellbeing, Mixed and Single Sex Groups

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	6,977	1	6,977	62,096	<,001	,272
	Greenhouse-Geisser	6,977	1,000	6,977	62,096	<,001	,272
	Huynh-Feldt	6,977	1,000	6,977	62,096	<,001	,272
	Lower-bound	6,977	1,000	6,977	62,096	<,001	,272
MeasuringPoints * typeofgroup.1	Sphericity Assumed	2,841	1	2,841	25,282	<,001	,132
	Greenhouse-Geisser	2,841	1,000	2,841	25,282	<,001	,132
	Huynh-Feldt	2,841	1,000	2,841	25,282	<,001	,132
	Lower-bound	2,841	1,000	2,841	25,282	<,001	,132
Error(MeasuringPoints)	Sphericity Assumed	18,652	166	,112			
	Greenhouse-Geisser	18,652	166,000	,112			
	Huynh-Feldt	18,652	166,000	,112			
	Lower-bound	18,652	166,000	,112			

Figure 15: Tests of Within-Subjects Effects, Psychosocial Wellbeing, Mixed and Single Sex Groups

We observe in Figure 16 that psychosocial wellbeing (on a scale from one to five) of boys and teenage boys increases to a larger degree within mixed teams than within same sex teams. Both Cohen's F (F = 0.390) and partial eta squared (partial η^2 = 0.132) show a large effect size.

Estimated Marginal Means of MEASURE_1
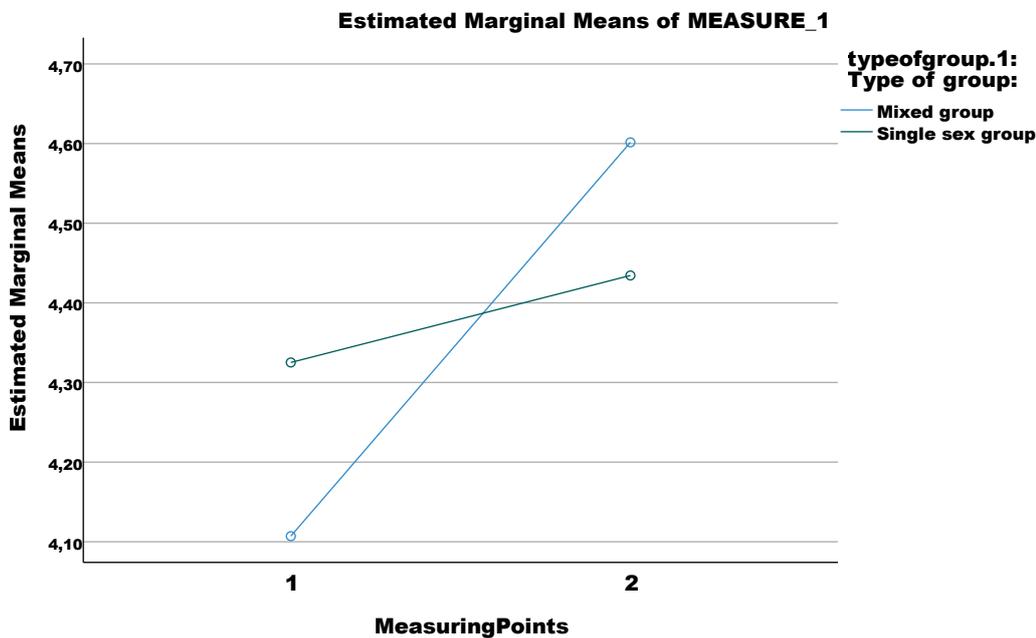
Figure 16: Estimated Marginal Means, Psychosocial Wellbeing, Mixed and Single Sex Groups

We thus confirm the hypothesis on male participants benefitting from girls' participation in sport teams and conclude that there is a significant interaction between the type of sport team and time, with male participants of mixed teams increasing their psychosocial wellbeing to a larger degree than male participants of same sex (all-boys) teams, Greenhouse-Geisser $F(1.00, 166.00) = 25.282$, $p < 0.001$, with a large effect size (Cohen's $F = 0.390$, partial $\eta^2 = 0.132$).

Mixed sport teams increase psychosocial wellbeing of boys more than all-boys teams.

Male participants benefit from the presence of female participants regarding atmosphere, social interactions, and group dynamics – all factors contributing to psychosocial wellbeing. In a context with such strict gender segregation this result shows the importance of questioning persistent gender norms in society.

Psychosocial Wellbeing in the IDP and Host Community Context

There were also differences found in psychosocial wellbeing depending on the background of the participants (IDP or host community).

Between-Subjects Factors

		Value Label	N
HostorCamp.1: Is the interviewee part of the...	1,00	Host community	35
	2,00	Camp	194

Figure 17: Between-Subjects Factors, Psychosocial Wellbeing, IDP Camps and Host Community

Descriptive Statistics

	HostorCamp.1: Is the interviewee part of the...	Mean	Std. Deviation	N
PsychosocialWellbeing.1	Host community	4,5625	,32971	35
	Camp	4,1463	,42639	194
	Total	4,2099	,43890	229
PsychosocialWellbeing.2	Host community	4,5911	,34686	35
	Camp	4,5302	,35052	194
	Total	4,5395	,34990	229

Figure 18: Descriptive Statistics, Psychosocial Wellbeing, IDP Camps and Host Community

Homogeneity of variances was asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.142) and in the endline (p = 0.760). Since it is an unbalanced design, this is an important perquisite.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
PsychosocialWellbeing.1	Based on Mean	2,278	1	227	,133
	Based on Median	2,171	1	227	,142
	Based on Median and with adjusted df	2,171	1	222,711	,142
	Based on trimmed mean	2,318	1	227	,129
PsychosocialWellbeing.2	Based on Mean	,136	1	227	,712
	Based on Median	,094	1	227	,760
	Based on Median and with adjusted df	,094	1	226,865	,760
	Based on trimmed mean	,172	1	227	,679

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + HostorCamp.1
   Within Subjects Design: MeasuringPoints

Figure 19: Levene's Test of Equality of Error Variances, Psychosocial Wellbeing, IDP Camps and Host Community

Box's Test of Equality of Covariance Matrices[a]

Box's M	3,405
F	1,110
df1	3
df2	47027,817
Sig.	,343

Figure 20: Box's Test of Equality of Covariance Matrices, Psychosocial Wellbeing, IDP Camps and Host Community

The same applies for homogeneity in covariance, which we can assume with Box's Test of Equality of Covariance Matrices showing p = 0.343. The assumption of sphericity can be neglected, as it only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

The mixed between-within ANOVA conducted to assess the influence of background on psychosocial wellbeing across two time periods (pre-intervention, post-intervention) shows a significant interaction between background and time (p < 0.001):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	2,522	1	2,522	20,592	<,001	,083
	Greenhouse-Geisser	2,522	1,000	2,522	20,592	<,001	,083
	Huynh-Feldt	2,522	1,000	2,522	20,592	<,001	,083
	Lower-bound	2,522	1,000	2,522	20,592	<,001	,083
MeasuringPoints * HostorCamp.1	Sphericity Assumed	1,871	1	1,871	15,274	<,001	,063
	Greenhouse-Geisser	1,871	1,000	1,871	15,274	<,001	,063
	Huynh-Feldt	1,871	1,000	1,871	15,274	<,001	,063
	Lower-bound	1,871	1,000	1,871	15,274	<,001	,063
Error(MeasuringPoints)	Sphericity Assumed	27,806	227	,122			
	Greenhouse-Geisser	27,806	227,000	,122			
	Huynh-Feldt	27,806	227,000	,122			
	Lower-bound	27,806	227,000	,122			

Figure 21: Tests of Within-Subjects Effects, Psychosocial Wellbeing, IDP Camps and Host Community

Figure 22 illustrates that psychosocial wellbeing (scale one to five) of participants of IDP camps is lower than the psychosocial wellbeing of members of the host community but can be increased to a larger degree through Sport for Development. Both Cohen's F (F = 0.260) and partial eta squared (partial η^2 = 0.063) show a medium effect size.

Estimated Marginal Means of MEASURE_1

HostorCamp.1: Is the interviewee part of the...
— Host community
— Camp
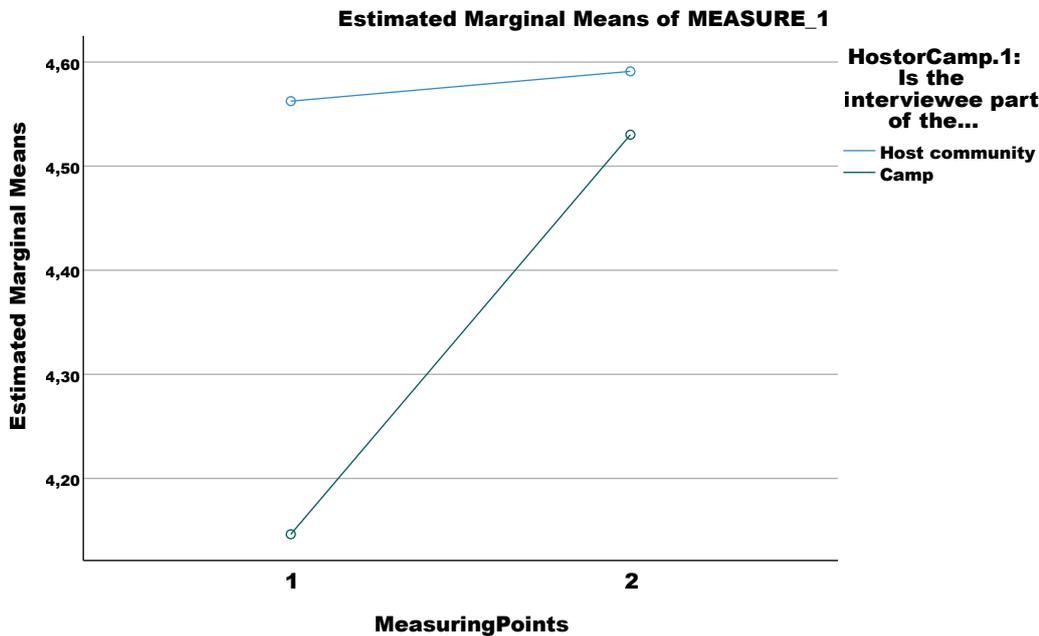
Figure 22: Estimated Marginal Means, Psychosocial Wellbeing, IDP Camps and Host Community

In conclusion, there is a significant interaction between the S4D participants' background and time, with participants of IDP camps increasing their psychosocial wellbeing to a larger degree than members of the host communities, Greenhouse-Geisser $F(1.00, 227.00) = 15.274$, $p < 0.001$, with a medium effect size (Cohen's $F = 0.260$, partial $\eta^2 = 0.063$).

> **S4D has a higher impact on psychosocial wellbeing among children and youth from IDP camps than from host communities.**

This result may not be surprising at first, as children and youth from IDP camps (have) experience(d) bigger hardships such as poverty, uncertain future, higher likelihood of having experienced traumatic events, poor living conditions, economic hardship, loss of related parties and torn families. Thus, there is a stronger lifting effect in children and youth from IDP camps to increase psychosocial wellbeing. While many international and local NGOs in the KRI refrain from working in IDP camps leading to a decrease of offers specifically for children and youth in the past years, this result is a strong reason for continuing to work in IDP camps.

Impacts on SDG Target 3.4: Social-Emotional Competences

To analyse S4D's impact on social-emotional competences, we created a scale with 14 Likert scale items measuring social-emotional competences revolving around six competences: self-confidence, trust, tolerance, communication, showing emotions and decision-making. As these competences do not necessarily mutually depend on each other we do not expect Cronbach's Alpha to show internal consistency of the scale. One person can have high self-confidence but low communication skills. However, we classify all six competences as the construct of social-emotional competences. Cronbach's Alpha in the baseline is 0.561 and in the endline 0.477 – confirming our expectations.

Reliability Statistics

Cronbach's Alpha	N of Items
,561	14

Figure 23: Cronbach's Alpha, Social-Emotional Competences, Baseline

Reliability Statistics

Cronbach's Alpha	N of Items
,477	14

Figure 24: Cronbach's Alpha, Social-Emotional Competences, Endline

To analyse social-emotional competences between the intervention and comparison group and within each group over time, a mixed between-within ANOVA was conducted. The normality assumption is neglected as the sample size is large and there is a balanced design.

Between-Subjects Factors

		Value Label	N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	229
	2,00	Control group (no S4D)	235

Figure 25: Between-Subjects Factors, Social-Emotional Competences

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
SocialEmotionalCompetences.1	Intervention group (S4D)	2,9314	,39655	229
	Control group (no S4D)	2,9506	,40592	235
	Total	2,9412	,40100	464
SocialEmotionalCompetences.2	Intervention group (S4D)	3,0661	,31338	229
	Control group (no S4D)	2,9294	,34317	235
	Total	2,9968	,33550	464

Figure 26: Descriptive Statistics, Social-Emotional Competences

Homogeneity of variances was asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.844) and in the endline (p = 0.642).

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
SocialEmotionalCompetences.1	Based on Mean	,023	1	462	,879
	Based on Median	,039	1	462	,844
	Based on Median and with adjusted df	,039	1	457,841	,844
	Based on trimmed mean	,030	1	462	,863
SocialEmotionalCompetences.2	Based on Mean	,201	1	462	,654
	Based on Median	,216	1	462	,642
	Based on Median and with adjusted df	,216	1	443,490	,642
	Based on trimmed mean	,218	1	462	,641

Figure 27: Levene's Test of Equality of Error Variances, Social-Emotional Competences

Box's Test of Equality of Covariance Matrices[a]

Box's M	2,702
F	,896
df1	3
df2	39249857,363
Sig.	,442

Figure 28: Box's Test of Equality of Covariance Matrices, Social-Emotional Competences

Homogeneity in covariance is assumed, with Box's Test of Equality of Covariance Matrices showing p = 0.442. The assumption of sphericity can be neglected, as it only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

The mixed between-within ANOVA conducted to assess the impact of Sport for Development on social-emotional competences across two time periods (pre-intervention, post-intervention) shows a significant interaction between S4D and time (p < 0.001):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	,746	1	,746	5,970	,015	,013
	Greenhouse-Geisser	,746	1,000	,746	5,970	,015	,013
	Huynh-Feldt	,746	1,000	,746	5,970	,015	,013
	Lower-bound	,746	1,000	,746	5,970	,015	,013
MeasuringPoints * InterControlGroup	Sphericity Assumed	1,410	1	1,410	11,286	<,001	,024
	Greenhouse-Geisser	1,410	1,000	1,410	11,286	<,001	,024
	Huynh-Feldt	1,410	1,000	1,410	11,286	<,001	,024
	Lower-bound	1,410	1,000	1,410	11,286	<,001	,024
Error(MeasuringPoints)	Sphericity Assumed	57,712	462	,125			
	Greenhouse-Geisser	57,712	462,000	,125			
	Huynh-Feldt	57,712	462,000	,125			
	Lower-bound	57,712	462,000	,125			

Figure 29: Tests of Within-Subjects Effects, Social-Emotional Competences

We observe in Figure 30 that social-emotional competences increase more within the intervention group than within the comparison group (which shows a decrease in the analysed competences), however the increase is rather small which is underlined by a small effect size: Cohen's F (F = 0.157) and partial eta squared (partial η^2 = 0.024).
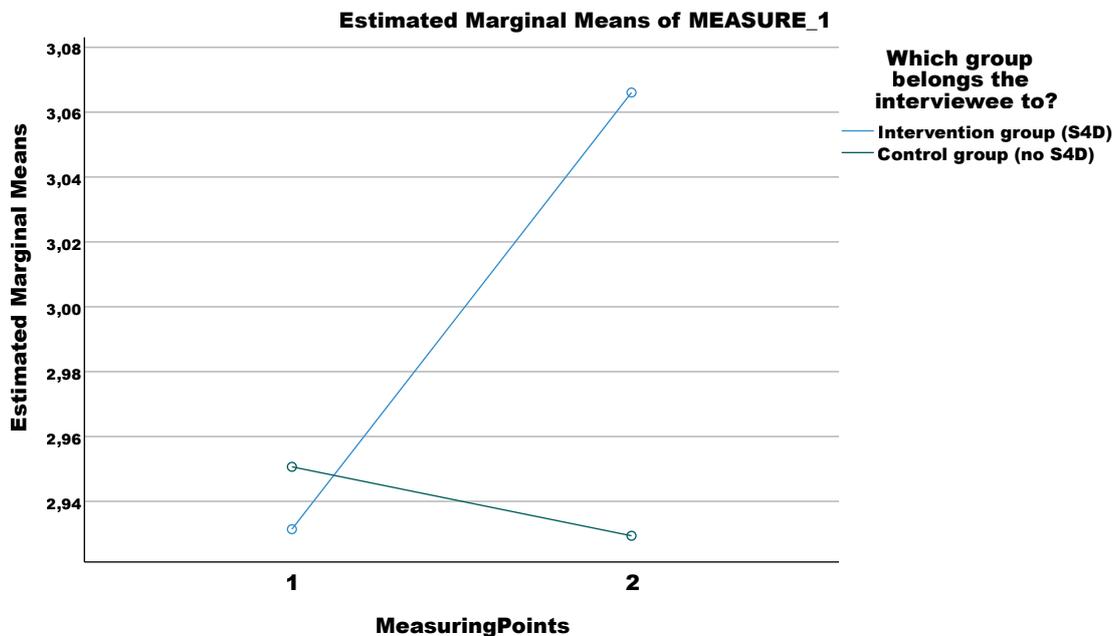
Figure 30: Estimated Marginal Means, Social-Emotional Competences

In conclusion, there is a significant interaction between S4D participation and time, with S4D participants increasing their social-emotional competences, Greenhouse-Geisser $F(1.00, 462.00)$ = 11.286, $p < 0.001$, with a small

S4D increases social-emotional competences to a small degree.

effect size (Cohen's F = 0.157, partial η^2 = 0.024). In psychological research, positive effects on social-emotional competences often turn out to be low, this is partly explained by the complexity of such competences which reflects our results (Ang et al., 2022; Liu, 2020).

Impact on SDG Target 5.1: Gender Equality

Measuring attitudes towards gender equality being a latent construct, we created a scale consisting of eight variables. Cronbach's Alpha shows good internal consistency of the scale for both baseline and endline with values of 0.658 (baseline) and 0.730 (endline).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,658	,681	8

Figure 31: Cronbach's Alpha, Gender Equality, Baseline

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,730	,752	8

Figure 32: Cronbach's Alpha, Gender Equality, Endline

To compare misogynistic and egalitarian attitudes between the intervention and comparison group and within each group over time, a mixed between-within ANOVA was conducted. The normality assumption is neglected as the sample size is large and it is a balanced design.

Between-Subjects Factors

		Value Label	N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	229
	2,00	Control group (no S4D)	235

Figure 33: Between-Subjects Factors, Gender Equality

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
GenderEquality.1	Intervention group (S4D)	3,2252	,55810	229
	Control group (no S4D)	3,2366	,58588	235
	Total	3,2309	,57175	464
GenderEquality.2	Intervention group (S4D)	3,7470	,32709	229
	Control group (no S4D)	3,2106	,61306	235
	Total	3,4753	,56099	464

Figure 34: Descriptive Statistics, Gender Equality

Homogeneity of variances was partly asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.334) but not in the endline (p < 0.001). Since variance homogeneity can be neglected with large sample sizes and balanced designs, this prerequisite can be neglected.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
GenderEquality.1	Based on Mean	1,105	1	462	,294
	Based on Median	,937	1	462	,334
	Based on Median and with adjusted df	,937	1	459,180	,334
	Based on trimmed mean	1,131	1	462	,288
GenderEquality.2	Based on Mean	79,049	1	462	<,001
	Based on Median	72,631	1	462	<,001
	Based on Median and with adjusted df	72,631	1	416,467	<,001
	Based on trimmed mean	74,175	1	462	<,001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + InterControlGroup
   Within Subjects Design: MeasuringPoints

Figure 35: Levene's Test of Equality of Error Variances, Gender Equality

Box's Test of Equality of Covariance Matrices[a]

Box's M	85,938
F	28,512
df1	3
df2	39249857,363
Sig.	<,001

Figure 36: Box's Test of Equality of Covariance Matrices, Gender Equality

The assumption of homogeneity of covariance is violated as $p < 0.001$. However, with a large sample size and a balanced design, this assumption can also be neglected. The assumption of sphericity can be neglected, as this only applies for procedures with measurement repetition with more than two stages. In the present case there are only two measurement points of time.

The mixed between-within ANOVA conducted to assess the impact of Sport for Development on attitudes towards gender equality across two time periods (pre-intervention, post-intervention) shows a significant interaction between S4D and time ($p < 0.001$):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	14,260	1	14,260	61,103	<,001	,117
	Greenhouse-Geisser	14,260	1,000	14,260	61,103	<,001	,117
	Huynh-Feldt	14,260	1,000	14,260	61,103	<,001	,117
	Lower-bound	14,260	1,000	14,260	61,103	<,001	,117
MeasuringPoints * InterControlGroup	Sphericity Assumed	17,400	1	17,400	74,558	<,001	,139
	Greenhouse-Geisser	17,400	1,000	17,400	74,558	<,001	,139
	Huynh-Feldt	17,400	1,000	17,400	74,558	<,001	,139
	Lower-bound	17,400	1,000	17,400	74,558	<,001	,139
Error(MeasuringPoints)	Sphericity Assumed	107,818	462	,233			
	Greenhouse-Geisser	107,818	462,000	,233			
	Huynh-Feldt	107,818	462,000	,233			
	Lower-bound	107,818	462,000	,233			

Figure 37: Tests of Within-Subjects Effects, Gender Equality

Figure 38 shows a scale from one to four, one being misogynistic attitudes and four being egalitarian attitudes. We observe a positive development within the intervention group over time and a slight decrease within the comparison group. The effect size is large with partial eta squared = 0.139 and Cohen's F = 0.402.

Estimated Marginal Means of MEASURE_1
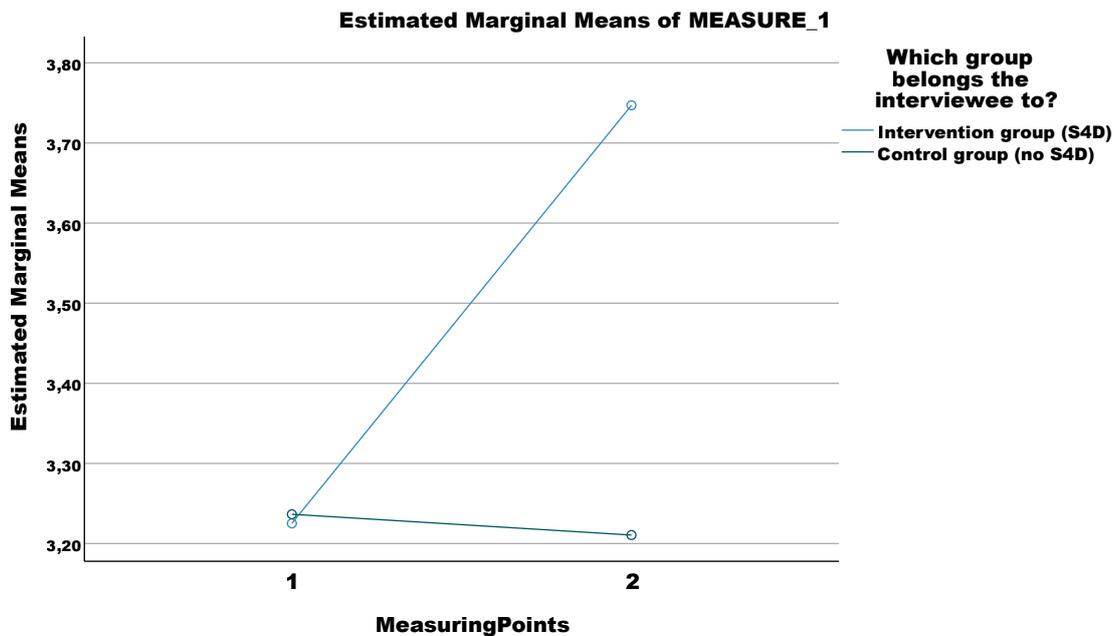
Figure 38: Estimated Marginal Means, Gender Equality

In conclusion, there is a significant interaction between S4D participation and time, with S4D contributing to egalitarian attitudes, Greenhouse-Geisser $F(1.00, 462.00) = 74.558$, $p < 0.001$, with a large effect size (Cohen's F = 0.402, partial $\eta^2 = 0.139$). That being the strongest and most meaningful result of this evaluation demonstrates how S4D strongly contributes to gender equality and empowerment of women and girls.

Results from the focus group discussions confirm this quantitative result: Especially girls in the older age groups show more awareness on gender equality and are demanding their rights while before, they deemed gender equality as "nothing to take seriously". Both quantitative and qualitative results show no significant differences between male and female attitudes. Discussions between male and female participants demonstrate their experience in debating issues revolving around gender equality in S4D trainings. While the quantitative results show no differences in attitudes towards gender equality based on age, in the focus group

S4D has a strong impact on SDG Target 5.1 and improves attitudes towards gender equality.

discussions some differences were observed: younger children of both genders tend to focus on physical limitations of women using this as a reason why girls and women cannot do physical or heavy labour that surpasses their alleged physical capacities. Older female participants understand how this kind of thinking is rooted deep into societal norms and traditions. This shows how essential safe spaces for critical reflection and discussion of such topics are, especially in a society that usually does not allow questioning certain norms. By providing such spaces, S4D teaches children and youth to discuss, listen, and critically reflect. One female participant puts it this way:

> "Before we participated in the S4D program, we thought it is not hard to be a woman in our community. Now, we know more about gender equality and our coaches made us aware about a lot of things and rights we did not even know about. Now we realized, there are a lot of rights we have to fight for as women."

In light of the recently launched feminist development policy of the German Federal Ministry for Economic Cooperation and Development, these results show how S4D can be used as a tool to fight misogynism. Through sports pedagogical approaches, gender roles, relations and norms are addressed and reflected. Girls and boys are sensitized to their rights as children and youth, and in particular to women's and human rights. The inclusion of boys and men plays an important role, as their questioning and breaking down of stereotypical gender norms is central to promoting positive masculinity, respectful gender relations and a peaceful society. Preconditions to achieve such results are safe spaces to discuss and reflect existing gender norms and the promotion of mutual respect, communication and teamwork through team sports and joint trainings. The results in the KRI demonstrate how S4D can transform constraining gender norms confirming S4D's gender transformative potential.

Impacts on SDG Target 5.2: Gender-based Violence against Women and Girls

We measure gender-based violence against women and girls (VAWG) by creating a scale of nine variables since gender-based VAWG is a latent construct. Cronbach's Alpha shows very good internal consistency of the scale for both baseline and endline with values of 0.749 (baseline) and 0.831 (endline).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,749	,749	9

Figure 39: Cronbach's Alpha, Gender-Based VAWG, Baseline

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,831	,831	9

Figure 40: Cronbach's Alpha, Gender-Based VAWG, Endline

To analyse perceptions and behavioural changes regarding gender-based VAWG between the intervention and comparison group and within each group over time, a mixed between-within ANOVA was conducted. The normality assumption is neglected as the sample size is large and it is a balanced design.

Between-Subjects Factors

	Value Label		N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	229
	2,00	Control group (no S4D)	234

Figure 41: Between-Subjects Factors, Gender-Based VAWG

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
GenderbasedViolence.1	Intervention group (S4D)	2,7729	,70803	229
	Control group (no S4D)	2,7799	,74005	234
	Total	2,7765	,72362	463
GenderbasedViolence.2	Intervention group (S4D)	3,2824	,67503	229
	Control group (no S4D)	2,6639	,77471	234
	Total	2,9698	,78957	463

Figure 42: Descriptive Statistics, Gender-Based VAWG

Homogeneity of variances was partly asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.409) but not in the endline (p < 0.004). Since variance homogeneity can be neglected with large sample sizes and balanced designs, this prerequisite is neglected.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
GenderbasedViolence.1	Based on Mean	,696	1	461	,405
	Based on Median	,683	1	461	,409
	Based on Median and with adjusted df	,683	1	460,466	,409
	Based on trimmed mean	,722	1	461	,396
GenderbasedViolence.2	Based on Mean	6,531	1	461	,011
	Based on Median	8,241	1	461	,004
	Based on Median and with adjusted df	8,241	1	446,572	,004
	Based on trimmed mean	7,994	1	461	,005

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + InterControlGroup
   Within Subjects Design: MeasuringPoints

Figure 43: Levene's Test of Equality of Error Variances, Gender-Based VAWG

Box's Test of Equality of Covariance Matrices[a]

Box's M	10,037
F	3,330
df1	3
df2	38826333,778
Sig.	,019

Checking for homogeneity in covariance in the case of a large sample, as recommended by Mertler (2004) the Box's Test is tested at a 0.001 significance level. With p = 0.019 homogeneity in covariance is asserted. The assumption of sphericity can be neglected, as this only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

Figure 44: Box's Test of Equality of Covariance Matrices, Gender-Based VAWG

The mixed between-within ANOVA conducted to assess the impact of S4D on perceptions and behavioural changes regarding gender-based VAWG across two time periods (pre-intervention, post-intervention) shows a significant interaction between S4D and time (p < 0.001):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	8,959	1	8,959	21,471	<,001	,045
	Greenhouse-Geisser	8,959	1,000	8,959	21,471	<,001	,045
	Huynh-Feldt	8,959	1,000	8,959	21,471	<,001	,045
	Lower-bound	8,959	1,000	8,959	21,471	<,001	,045
MeasuringPoints * InterControlGroup	Sphericity Assumed	22,651	1	22,651	54,285	<,001	,105
	Greenhouse-Geisser	22,651	1,000	22,651	54,285	<,001	,105
	Huynh-Feldt	22,651	1,000	22,651	54,285	<,001	,105
	Lower-bound	22,651	1,000	22,651	54,285	<,001	,105
Error(MeasuringPoints)	Sphericity Assumed	192,356	461	,417			
	Greenhouse-Geisser	192,356	461,000	,417			
	Huynh-Feldt	192,356	461,000	,417			
	Lower-bound	192,356	461,000	,417			

Figure 45: Tests of Within-Subjects Effects, Gender-Based VAWG

Figure 46 shows the acceptance of gender-based VAWG on a scale from one to four, one being misogynistic attitudes and four being egalitarian attitudes. We observe a positive development within the intervention group over time and a slight decrease within the comparison group. The effect size is medium to large with partial eta squared = 0.105 and Cohen's F = 0.343.

Estimated Marginal Means of MEASURE_1
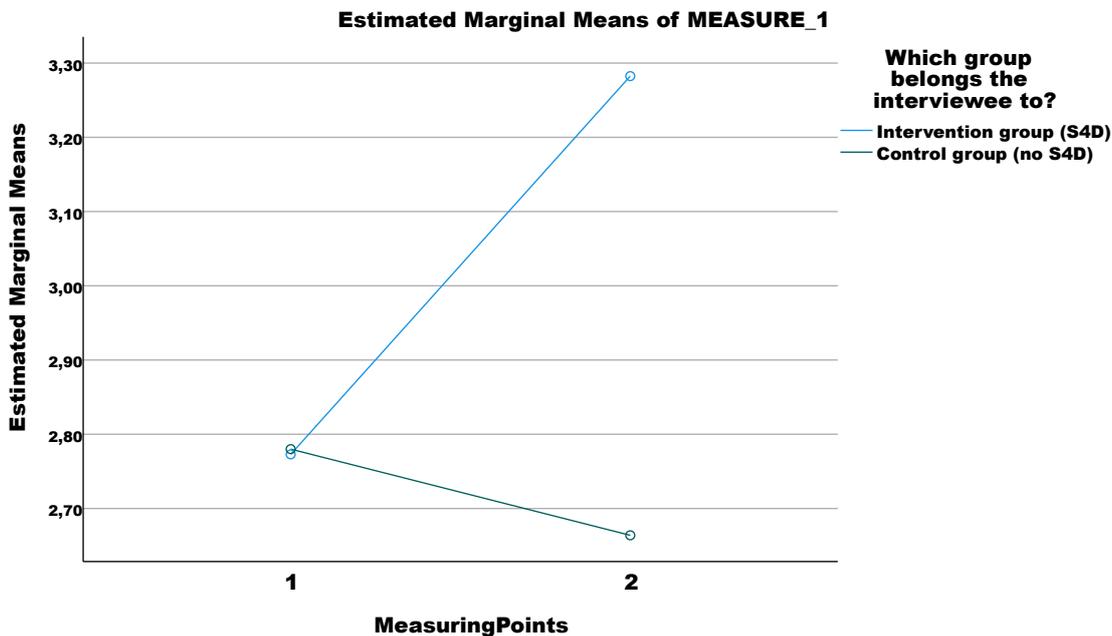
Figure 46: Estimated Marginal Means, Gender-Based VAWG

In conclusion, there is a significant interaction between S4D participation and time, with S4D reducing the acceptance of gender-based VAWG, Greenhouse-Geisser $F(1.00, 461.00) = 54.285$, $p < 0.001$, with a medium to large effect size (Cohen's $F = 0.343$, partial $\eta^2 = 0.105$).

In the context of crises and conflicts, misogynist attitudes are often intensifying and sexualized – and gender-based violence against women and girls is on the rise. The S4D intervention in the KRI was designed as a non-discriminatory place that acts preventively against gender-based VAWG. Raising awareness of women's and human rights played a central role in this, as did the involvement of boys and men in order to question and break down

> **S4D reduces the acceptance of and propensity for gender-based violence against women and girls and has a strong impact on SDG Target 5.2.**

constraining gender norms and toxic masculinity. Results from the focus group discussions underline this. Women and girls tend to withdraw from many social activities out of fear of being sexually harassed. For many of the participants, S4D trainings were the first-time doing sports with the other gender. S4D trainings thus provided a space for social

interaction between male and female participants and to commonly learn about a topic that affects male and female participants in two extremes: one in the role of a possible offender and one in the role of a possible victim. Discussions between participants show an ongoing reflection process about gender-based VAWG and their own roles, rights, and the law. Once again this demonstrates the need for safe spaces for reflection, critical discussion, and learning.

Addressing such topics can put female participants in a vulnerable position. That is why safe spaces for girls were created and girls had the option to choose between mixed teams and all-girls teams. Additionally, girls-only trainings were offered for girls participating in mixed teams. Strongly interlinked with SDG Target 5.1, gender-based VAWG can be successfully addressed in S4D implementations and as this study confirms, S4D can reduce the acceptance and thus the propensity for gender-based violence against women and girls – making S4D a meaningful tool also for development projects working on gender equality and the reduction of VAWG.

Impact on SDG Target 5.4:
Valuing Unpaid Work and Domestic Care

In addition to impacts on SDG Target 5.1 and SDG Target 5.2, we checked for changes in perceptions of unpaid work and domestic care which, in the Kurdistan Region of Iraq, is traditionally a responsibility performed by women and girls. Promoting shared responsibility within households and families and valuing such care work requires a holistic approach that targets not only children and youth but families, communities as a whole, and other essential gatekeepers. However, it is of interest to evaluate if and to what extent S4D can contribute to SDG Target 5.4 as an unintended result of S4D trainings targeting gender equality.

We analysed responsibilities within the participants' households, to obtain a clearer picture of the prevailing domestic roles:

		HouseholdLaborDivision.1: In my community, girls and women are responsible for the household.				
		I totally agree	I somewhat agree	I somewhat disagree	I totally disagree	Total
Which group belongs the interviewee to?	Intervention group (S4D)	201	35	12	54	302
	Control group (no S4D)	221	27	7	43	298
Total		422	62	19	97	600

Figure 47: Household Responsibilities, Baseline

		HouseholdLaborDivision.2: In my community, girls and women are ...				
		I totally agree	I somewhat agree	I somewhat disagree	I totally disagree	Total
Which group belongs the interviewee to?	Intervention group (S4D)	146	34	9	46	235
	Control group (no S4D)	182	29	10	31	252
Total		328	63	19	77	487

Figure 48: Household Responsibilities, Endline

As the two tables show, both in the baseline as well as the endline, data depicts a rather traditional division of roles and responsibilities within households – for both intervention and comparison group and over both measuring points (baseline and endline).

In the baseline, 78.2% of the intervention group and 83.3% of the comparison group state that household responsibilities at home lie with women and girls. In the endline, 76.6% of the intervention group and 83.7% of the comparison group state that household responsibilities at home lie with women and girls. Focus group discussions underline these figures. "Of course, men should help with chores" was a common statement. While this topic triggered some deep discussions between female and male participants, there always came a big "but" with manyfold justifications for a traditional labour division in households: Lack of time among men; males as main bread winner; women being more capable and willing to perform such household chores; division of chores into more "male" (technical) and "female" tasks and the perception of "never change a running system". This shows that the concept of gender equality is not lived or internalized in all aspects of daily life. While women may work, generate money and study – they still remain the main person responsible for household chores and care work.

We analysed possible changes in perceptions on traditional household roles between the intervention and comparison group and within each group over time by conducting a mixed between-within ANOVA. The normality assumption can be neglected as the sample size is large and it is a balanced design.

Between-Subjects Factors

		Value Label	N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	222
	2,00	Control group (no S4D)	231

Figure 49: Between-Subjects Factors, Unpaid Domestic Work

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
UnpaidWork.1	Intervention group (S4D)	2,4369	1,38613	222
	Control group (no S4D)	2,5368	1,41027	231
	Total	2,4879	1,39784	453
UnpaidWork.2	Intervention group (S4D)	3,4234	,96124	222
	Control group (no S4D)	2,6320	1,35410	231
	Total	3,0199	1,24163	453

Figure 50: Descriptive Statistics, Unpaid Domestic Work

Homogeneity of variances was partly asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.762) but not in the endline (p < 0.001). Since variance homogeneity can be neglected with large sample sizes and balanced designs, this prerequisite can be neglected.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
UnpaidWork.1	Based on Mean	,906	1	451	,342
	Based on Median	,092	1	451	,762
	Based on Median and with adjusted df	,092	1	449,742	,762
	Based on trimmed mean	,901	1	451	,343
UnpaidWork.2	Based on Mean	114,747	1	451	<,001
	Based on Median	72,162	1	451	<,001
	Based on Median and with adjusted df	72,162	1	400,142	<,001
	Based on trimmed mean	115,007	1	451	<,001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + InterControlGroup
Within Subjects Design: MeasuringPoints

Figure 51: Levene's Test of Equality of Error Variances, Unpaid Domestic Work

Box's Test of Equality of Covariance Matrices[a]

Box's M	26,012
F	8,629
df1	3
df2	38536881,595
Sig.	<,001

We check for homogeneity in covariance by using Box's Test of Equality of Covariance Matrices. In this case, the assumption of homogeneity of covariance is violated as $p < 0.001$. However, with a large sample size and a balanced design, this assumption can be neglected.

The assumption of sphericity can be neglected, as this only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

Figure 52: Box's Test of Equality of Covariance Matrices, Unpaid Domestic Work

The mixed between-within ANOVA conducted to assess the impact of S4D on perceptions of unpaid work and domestic care across two time periods (pre-intervention, post-intervention) shows a significant interaction between S4D and time ($p < 0.001$):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	66,232	1	66,232	48,301	<,001	,097
	Greenhouse-Geisser	66,232	1,000	66,232	48,301	<,001	,097
	Huynh-Feldt	66,232	1,000	66,232	48,301	<,001	,097
	Lower-bound	66,232	1,000	66,232	48,301	<,001	,097
MeasuringPoints * InterControlGroup	Sphericity Assumed	44,961	1	44,961	32,788	<,001	,068
	Greenhouse-Geisser	44,961	1,000	44,961	32,788	<,001	,068
	Huynh-Feldt	44,961	1,000	44,961	32,788	<,001	,068
	Lower-bound	44,961	1,000	44,961	32,788	<,001	,068
Error(MeasuringPoints)	Sphericity Assumed	618,432	451	1,371			
	Greenhouse-Geisser	618,432	451,000	1,371			
	Huynh-Feldt	618,432	451,000	1,371			
	Lower-bound	618,432	451,000	1,371			

Figure 53: Tests of Within-Subjects Effects, Unpaid Domestic Work

Figure 54 shows misogynistic to egalitarian perceptions on shared household responsibilities on a scale from one to four, one being misogynistic attitudes and four being egalitarian attitudes. We observe a positive development within the intervention group over time and only a slight increase within the comparison group. The effect size is medium with partial eta squared = 0.068 and Cohen's F = 0.270.
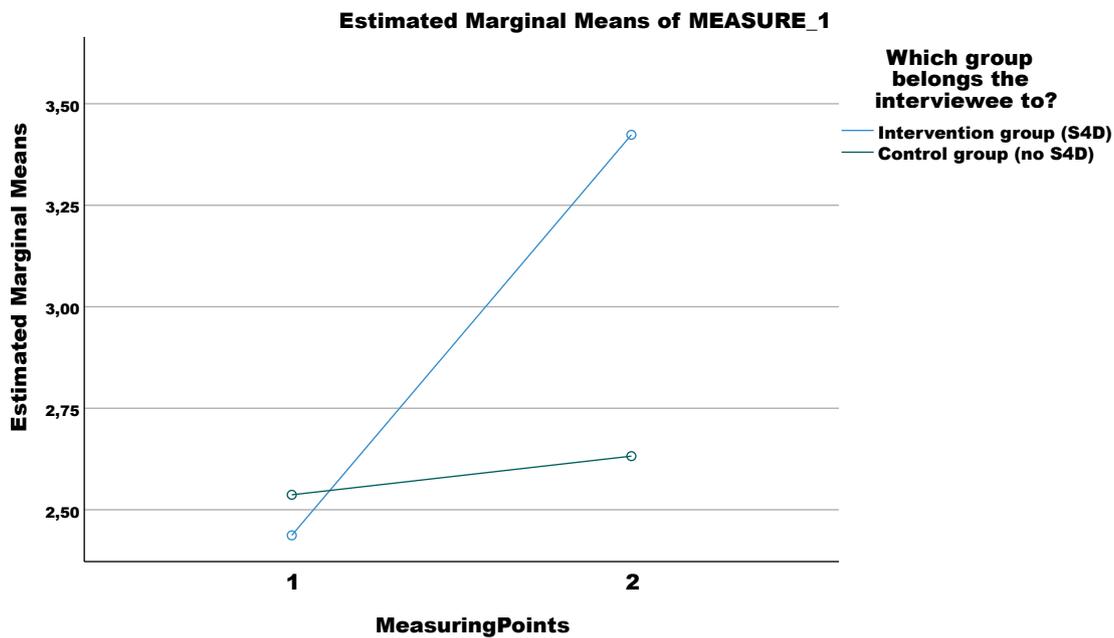
Figure 54: Estimated Marginal Means, Unpaid Domestic Work

In conclusion, there is a significant interaction between S4D participation and time, with S4D improving perceptions on equal division of household responsibilities, Greenhouse-Geisser $F(1.00, 451.00) = 44.961$, $p < 0.001$, with

> **S4D promotes attitudes towards an equal division of labour in the household but not behavioural changes**.

a medium effect size (Cohen's $F = 0.270$, partial $\eta^2 = 0.068$). We thus demonstrate an impact of S4D on the change of perceptions – however we cannot confirm any behavioural changes among participants. And such behavtioural changes are not expected as they need a more holistic and long-term project approach that S4D alone cannot fulfil. However, S4D can initiate changes in perception and trigger thought processes which demonstrates that S4D can be used as a door opener and an instrument that can be integrated into gender-transformative program approaches.

Impacts on SDG Target 10.2: Social Cohesion

While the S4D intervention in the KRI did not specifically target social cohesion, inclusion and the reduction of inequalities, the intervention still takes place in a multi-ethnic and multireligious context – thus we checked for unintended impacts on social cohesion.

We measure social cohesion by creating a scale of five variables since social cohesion is a latent construct. Cronbach's Alpha shows an acceptable internal consistency of the scale for both baseline and endline with values of 0.455 (baseline) and 0.636 (endline).

Reliability Statistics

Cronbach's Alpha	N of Items
,455	5

Figure 55: Cronbach's Alpha, Social Cohesion, Baseline

Reliability Statistics

Cronbach's Alpha	N of Items
,636	5

Figure 56: Cronbach's Alpha, Social Cohesion, Endline

It is assumed that the reason for a lower Cronbach's Alpha in the baseline is that questions on the topic of social cohesion were a new concept for participants and was only fully grasped in the endline after learning and discussing it in S4D trainings. To analyse social cohesion between the intervention and comparison group and within each group over time, a mixed between-within ANOVA was conducted. The normality assumption is neglected as the sample size is large and it has a balanced design.

Between-Subjects Factors

		Value Label	N
Which group belongs the interviewee to?	1,00	Intervention group (S4D)	229
	2,00	Control group (no S4D)	235

Figure 57: Between-Subjects Factors, Social Cohesion

Descriptive Statistics

	Which group belongs the interviewee to?	Mean	Std. Deviation	N
SocialCohesion.1	Intervention group (S4D)	3,2352	,68880	229
	Control group (no S4D)	3,2449	,70626	235
	Total	3,2401	,69696	464
SocialCohesion.2	Intervention group (S4D)	3,7721	,34696	229
	Control group (no S4D)	3,3606	,59809	235
	Total	3,5636	,53149	464

Figure 58: Descriptive Statistics, Social Cohesion

Homogeneity of variances was partly asserted using Levene's Test based on median which shows that equal variances can be assumed in the baseline (p = 0.990) but not in the endline (p < 0.001). Since variance homogeneity can be neglected with large sample sizes and balanced designs, this prerequisite can be neglected.

Levene's Test of Equality of Error Variances[a]

		Levene Statistic	df1	df2	Sig.
SocialCohesion.1	Based on Mean	,270	1	462	,604
	Based on Median	,000	1	462	,990
	Based on Median and with adjusted df	,000	1	461,225	,990
	Based on trimmed mean	,071	1	462	,791
SocialCohesion.2	Based on Mean	36,189	1	462	<,001
	Based on Median	33,668	1	462	<,001
	Based on Median and with adjusted df	33,668	1	384,924	<,001
	Based on trimmed mean	33,466	1	462	<,001

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + InterControlGroup
   Within Subjects Design: MeasuringPoints

Figure 59: Levene's Test of Equality of Error Variances, Social Cohesion

Box's Test of Equality of Covariance Matrices[a]

Box's M	65,931
F	21,874
df1	3
df2	39249857,363
Sig.	<,001

The assumption of homogeneity of covariance is violated as p < 0.001. However, with a large sample size and a balanced design this assumption can be neglected. The assumption of sphericity can be neglected, as this only applies for procedures with measurement repetition that have more than two stages. In the present case there are only two measurement points of time.

Figure 60: Box's Test of Equality of Covariance Matrices, Social Cohesion

The mixed between-within ANOVA conducted to assess the impact of S4D on social cohesion across two time periods (pre-intervention, post-intervention) shows a significant interaction between S4D and time (p < 0.001):

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
MeasuringPoints	Sphericity Assumed	24,690	1	24,690	81,200	<,001	,149
	Greenhouse-Geisser	24,690	1,000	24,690	81,200	<,001	,149
	Huynh-Feldt	24,690	1,000	24,690	81,200	<,001	,149
	Lower-bound	24,690	1,000	24,690	81,200	<,001	,149
MeasuringPoints * InterControlGroup	Sphericity Assumed	10,286	1	10,286	33,828	<,001	,068
	Greenhouse-Geisser	10,286	1,000	10,286	33,828	<,001	,068
	Huynh-Feldt	10,286	1,000	10,286	33,828	<,001	,068
	Lower-bound	10,286	1,000	10,286	33,828	<,001	,068
Error(MeasuringPoints)	Sphericity Assumed	140,477	462	,304			
	Greenhouse-Geisser	140,477	462,000	,304			
	Huynh-Feldt	140,477	462,000	,304			
	Lower-bound	140,477	462,000	,304			

Figure 61: Tests of Within-Subjects Effects, Social Cohesion

Figure 62 depicts exclusionary to inclusive behaviour on a scale from one to four, one being exclusionary behaviour and four being inclusive behaviour. We observe a positive development within the intervention group over time and only a slight increase within the comparison group. The effect size is medium with partial eta squared = 0.068 and Cohen's F = 0.270.
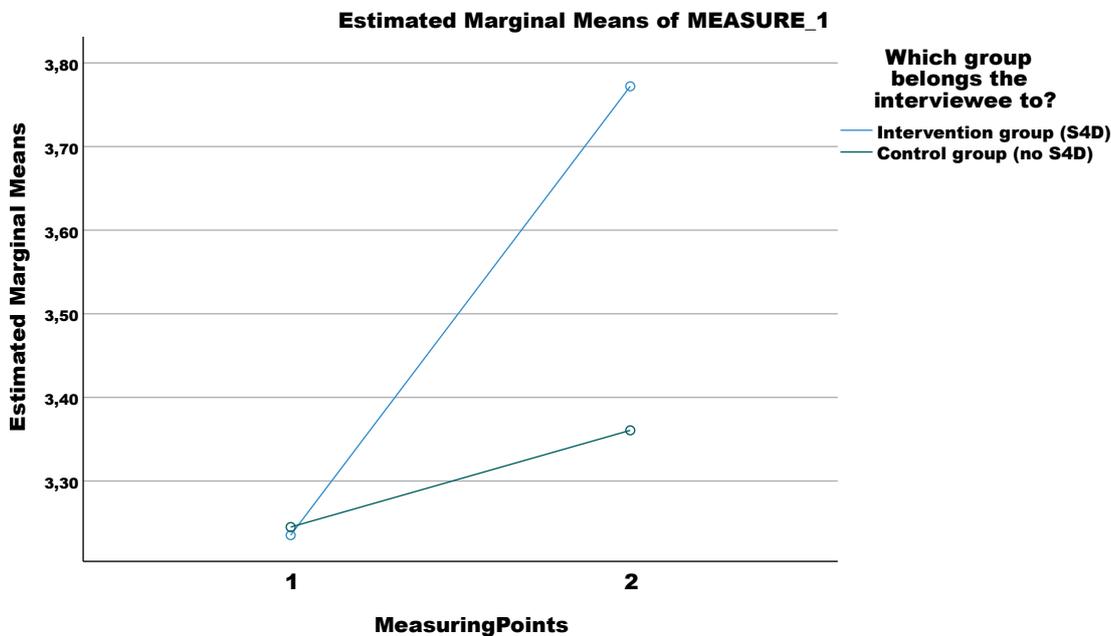
Figure 62: Estimated Marginal Means, Social Cohesion

In conclusion, there is a significant interaction between S4D participation and time, with S4D improving social cohesion, Greenhouse-Geisser $F(1.00, 4621.00) = 33.828$, $p < 0.001$, with a medium effect size (Cohen's $F = 0.270$, partial $\eta^2 = 0.068$).

S4D reinforces inclusive behaviour towards other religions and ethnicities and has an impact on SDG Target 10.2.

Sport in general and S4D in specific provide a framework in which topics of social cohesion and the breaking down of stereotypes can be addressed. Especially when sport activities are designed to encourage participation, essential competences such as critical thinking, respect, tolerance, cooperation and problem-solving can develop. Furthermore, S4D activities require interaction and can thus help challenge stereotypes and break barriers of mistrust between young people. By experiencing the realities of others, S4D can serve as a connecting element. The results in the KRI demonstrate the inclusive character of S4D and show that S4D contributes to inclusive behaviour and social cohesion.

LIMITATIONS OF THE STUDY

With the choice of a quasi-experimental, longitudinal study design instead of a Randomized Controlled Trial (RCT) we cannot exclude the possibility of a selection effect and cannot control for all disruptive factors distorting an unbiased assessment. However, working with a vulnerable group like children and youth from IDP camps, probability sampling would have provoked ethical concerns as outlined before. We acknowledge this limitation to our study design and took measures to control for it by choosing two points of measurement and ensuring consistency in the S4D implementation. By doing so, Stockmann (2007) argues that there are hardly any differences to a RCT in terms of design quality.

While we have a balanced design regarding intervention and comparison group, it would have been preferred to have a more balanced design with regards to subgroups, especially gender, type of sport (ultimate frisbee or football) and type of sport team (mixed or same sex team). With a gender ratio of 28% females in the intervention group and 37% in the comparison group, we are still rather far from an equal participation and representation of all genders. While the lower participation rate of females is due to local circumstances, more time and more awareness raising prior to the implementation might have led to a more balanced participation rate. The lower participation rate of females is also one reason for only one all-girls team (the other reason being the preference for mixed teams by females) making it impossible to compare psychosocial wellbeing of girls in same sex teams to the wellbeing of girls in mixed teams – as could be analysed for boys. In general, a more balanced design within the subgroups might have shown even more interesting results while in the present case, many of these analyses evaluating effects of S4D on different genders, age groups, sport teams, sport types and backgrounds were insignificant. With ultimate frisbee being a very new sport in the KRI, there was not a sufficient number of coaches available to ensure an equal amount of ultimate frisbee and football teams – nevertheless it might be interesting in the future to look further into the effects of different sport types.

Conducting questionnaires with children and youth always raises the issue of social desirability. We noticed a tendency towards "better answers", high approval rates and higher values in the response behaviour of the participants. This is a well-known phenomenon in social sciences and psychology. Due to social desirability, respondents and especially children and youth try to give a predominantly positive description of one's own person and to correspond to what the interviewer or other persons involved supposedly expect from them. This can be done by means of an exaggerated mention of desirable behaviour or by means of an understated mention of undesirable behaviour. Orientation is provided by social norms (Kreuter, F. et al. 2008). Additionally, children and youth often reflect their response behaviour differently after an intervention, knowing more about the different topics which sometimes even leads to supposedly negative results. While this is not the case in the present study, we identified the phenomenon of social desirability. We control for it by having a comparison group and by using quantitative analysis methods that are able to still identify significant results and sufficient effect sizes.

RECOMMENDATIONS

The success of the evaluated S4D implementation is due to three core elements which we recommend future S4D implementations to take into consideration: 1) An intervention period of minimum six months although we see this as a minimum and recommend interventions up to one year. Trainings twice per week increase the intensity of the intervention. 2) Intensive training and ongoing support of coaches: the coaches were accompanied during the whole implementation through a designated S4D instructor who monitored the trainings and schedules, provided feedback and psychological support to coaches when needed and served as a liaison between coaches, the implementing partner NGO and the GIZ project. 3) The pairing of coaches with different competences in trauma pedagogy and physical education. It was crucial to work with coaches from the same context as the target groups: Eight coaches are IDPs, eleven coaches come from the host community. At the same time, it was essential to create spaces for peer support, as it is a challenging working context for a partially affected and potentially traumatized person.

Success factors of the S4D intervention:

1) **Length and intensity of intervention**
2) **Qualified coaches, ongoing support structure and monitoring**
3) **Pairing of coaches**

For the intervention group, we recommend a more thorough mixing of IDPs and members of the host community which is specifically important with regards to social cohesion. In some cases, stricter adherence to the three specified age groups within teams (10-12, 13-15, 16-19) would have been needed, as younger and older age groups have different needs and learning processes. And as outlined above, an equal participation and representation of all genders is essential.

One prerequisite for implementors – individuals as well as organisations – working with mixed sport teams in a cultural context such as the KRI is the flexibility to adapt implementation; deep knowledge of the present context; trust from participants and families; and an overall sensitivity for the context and a vulnerable target group. By either ignoring or complying with this, projects either fail or succeed.

CONCLUSION

Results show that Sport for Development contributes to the achievement of SDG Target 3.4, SDG Target 5.1, SDG Target 5.2, SDG Target 5.4 and SDG Target 10.2. These results can be generalized to the studied context, thus to the Kurdistan Region of Iraq and the target group of 10 to 19 year old children and youth.

The study confirms that S4D measures have a statistically significant, medium effect on psychosocial wellbeing of children and youth and thus increase psychosocial wellbeing. We thus demonstrate that S4D can be used as an effective psychosocial support measure that fosters a sense of belonging, hears children's and youth's voices and right to be heard. Through S4D, children and youth experience self-efficacy, which is specifically important in the context of forced migration and potential traumatisation where stability, belonging and a healthy environment are very often lacking. This makes S4D an effective and meaningful approach not only for sport projects but also for development interventions working on MHPSS – in crisis or post crisis settings.

S4D also shows a statistically significant, albeit weak effect on social-emotional skills. In psychological research, positive effects on social-emotional competences often turn out to be low, this is partly explained by the complexity of such competences which reflects our results.

Mixed sport teams increase the psychosocial wellbeing of boys more compared to all-boys teams. Male participants benefit from the presence of female participants regarding atmosphere, social interactions, and group dynamics – all factors contributing to psychosocial wellbeing. In a context with such strict gender segregation this result shows the importance of questioning persistent gender norms in society.

The study also shows that the psychosocial wellbeing of children and youth in IDP camps is lower than that of children and youth from host communities. However, S4D increases the wellbeing of the former more than the latter. This result may not be surprising at first, as children and youth from IDP camps (have) experience(d) bigger hardships such as poverty, uncertain future, higher likelihood of having experienced traumatic events, poor living conditions, economic hardship, loss of related parties and torn families. Thus, there is a stronger lifting effect in children and youth from IDP camps to increase psychosocial wellbeing. As many international and local NGOs in the KRI refrain from working in IDP camps leading to a decrease of offers specifically for children and youth in the past years, this result is a strong reason for continuing to work in IDP camps.

S4D has a statistically significant, strong effect on SDG Target 5.1 and improves attitudes towards gender equality in the studied context. That being the strongest and most meaningful result of this evaluation and in light of the recently launched feminist development policy, this result shows how S4D can be used as a tool to fight misogynism. Through sport pedagogical approaches, gender roles, relations and norms are addressed and reflected. Girls and boys are sensitized to their rights as children and youth, and in particular to women's and human rights. The inclusion of boys and men plays an important role, as their questioning and breaking down of stereotypical gender norms is central to promoting positive masculinity, respectful gender relations and a peaceful society. Preconditions to achieve such results are safe spaces to discuss and reflect existing gender norms and the

promotion of mutual respect, communication and teamwork through team sports and joint trainings. The results in the KRI demonstrate how S4D can transform constraining gender norms confirming S4D's gender transformative potential.

S4D has a statistically significant, medium to strong effect on SDG Target 5.2 and reduces the propensity for gender-based violence against women and girls. In the context of crises and conflicts, misogynist attitudes are often intensifying and sexualized – and gender-based violence against women and girls is on the rise. The S4D intervention in the KRI was designed as a non-discriminatory place that acts preventively against gender-based VAWG. Raising awareness of women's and human rights played a central role in this, as did the involvement of boys and men in order to question and break down constraining gender norms and toxic masculinity. Safe spaces for girls were created and girls had the option to choose between mixed teams and all-girls teams. Additionally, girls-only trainings were offered for girls participating in mixed teams. Strongly interlinked with SDG Target 5.1, gender-based VAWG can be successfully addressed in S4D implementations and as this study confirms, S4D can reduce the acceptance and thus the propensity for gender-based violence against women and girls – making S4D an effective tool also for development projects working on gender equality and the reduction of VAWG.

The analyses also show a statistically significant, medium effect on SDG Target 5.4: S4D measures promote attitudes towards an equal division of labour in the household – but do not contribute to behavioural changes in this regard. The breaking down of social norms towards an equal division of labour requires a holistic and more long-term approach, which S4D alone cannot realize. However, S4D can initiate changes in perception and trigger thought processes which demonstrates that S4D can be used as a door opener and an instrument that can be integrated into gender-transformative program approaches.

The study also examined the contribution of S4D to SDG Target 10.2. A statistically significant, medium effect could be demonstrated: S4D strengthens inclusive behaviour towards other religions and ethnicities and promotes social cohesion. S4D provides a framework in which topics of social cohesion and the breaking down of stereotypes can be addressed. In a context that encourages participation, not only competition, essential competences such as critical thinking, respect, tolerance, cooperation and problem-solving can develop. S4D activities require interaction and can thus help challenge stereotypes and break barriers of mistrust by young people. By experiencing the realities of others, S4D can serve as a connecting element. The results in the KRI demonstrate the inclusive character of S4D and show that S4D contributes to inclusive behaviour and social cohesion.

These very positive results are linked to a very well thought out implementation logic, well adapted to the local context and target group. The sought-after competence acquisition, behavioural and attitude changes in the target group could only be achieved through a well-developed training manual and subsequent trainings, experienced and qualified coaches, continuous monitoring of the implementation, a sufficient S4D intervention period and trusted implementation partners. While short term or singular S4D measures might create the needed publicity for Sport for Development and can show results on output level, it is the longer term, intensive S4D interventions that create impact. As this study demonstrates.

LITERATURE

Ang, W. H. D., Lau, S. T., Cheng, L. J., Chew, H. S. J., Tan, J. H., Shorey, S., & Lau, Y. (2022). Effectiveness of resilience interventions for higher education students: A meta-analysis and metaregression. *Journal of Educational Psychology, 114(7)*, 1670–1694.

Fiedler, C. & Rohles, C. (2021). *Social Cohesion after Armed Conflict: A Literature Review.* Discussion Paper. Bonn: German Development Institute / Deutsches Institut für Entwicklungspolitik (DIE).

GIZ (2022). *Sports Builds Gender Equality. A practical guide to foster gender equality and prevent gender-based violence against women and girls through football*. Retrieved on 25.04.2023 from https://www.sport-for-development.com/imglib/downloads/Manuale/Iraq/giz2022-en-manual-sport%20builds-gender-equality-iraq.pdf.

Kreuter, F., Presser, S, & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys. P*ublic Opinion Quarterly*, 72(5), 847-865.

Leininger, J., von Schiller, A., Sommer, C., Burchi, F., Fiedler, C., Mross, K., Nowack, D., & Strupat, C. (2020). *Background note: Social cohesion – Concept, measurement and country profiles* (unpublished Background Note). Bonn: German Development Institute / Deutsches Institut für Entwicklungspolitik (DIE).

Liu, J. J. W., Ein, N., Gervasio, J., Battaion, M., Reed, M., & Vickers, K. (2020). Comprehensive meta-analysis of resilience in terventions. *Clinical Psychology Review*, 82 (101919).

Mertler, C. A. (2004). *Advanced and Multivariate Statistical Methods: Practical Application and Interpretation* (3rd ed.). Pyrczak Publishing.

Ravens-Sieberer, U., Herdman, M., Devine, J., Otto, C., Bullinger, M., Rose, M., & Klasen, F. (2014). The European KIDSCREEN approach to measure quality of life and well-being in children: Development, current application, and future advances. *Quality of Life Research*, 23(3), 791-803.

REACH (2021). *Iraq IDP Camp Profiling. Round XV. Camp Directory*. June-August 2021. Retrieved on 28.03.2023 from Iraq: IDP Camp Profiling, Camp Directory - Round XV, June-August 2021 - Iraq | ReliefWeb.

Salkind, N. J. (2010). *Encyclopedia of Research Design* (Vol. 2). Los Angeles: Sage.

Stockmann, R. (2007). *Handbuch zur Evaluation. Eine praktische Handlungsanleitung*. Münster: Waxmann Verlag GmbH.

Supreme Judicial Court of Iraq (2022): *More than 18,000 cases of rape against women*. Retrieved on 08.05.2023 from https://www.sjc.iq/view.70122/.

Tabachnik, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Pearson Education.

UNHCR (2022). I*raq Camp Master List and Population Flow - November 2022*. Retrieved on 28.03.2023 from https://data.unhcr.org/en/documents/details/97463.

UN WOMEN, MADRE, City University of New York (2021): I*dentifying Gender Persecution in Conflict and Atrocities. A Toolkit for Documenters, Investigators, Prosecutors and Adjudicators of Crimes against Humanity*. Retrieved on 28.04.2023 from https://www.unwomen.org/sites/default/files/2022-01/Identifying-gender-persecution-in-conflict-and-atrocities-en.pdf.

Verma, J. P. (2015). *Repeated Measures Design for Empirical Researchers* (1st ed.). Wiley.

Warner, R. M. (2012). *Applied Statistics: From Bivariate Through Multivariate Techniques* (2nd ed.). Sage Publications Ltd.

World Health Organisation (2019). *Mental Health Atlas 2017*. Retrieved on 19.02.2023 from https://www.who.int/publications/i/item/mental-health-atlas-2017.