2. METHOD

This dissertation study was part of the research project "Interactive Brains, Social Minds" at the Center for Lifespan Psychology, Max Planck Institute (MPI) for Human Development in Berlin, Germany. The aim of this project is to investigate the development of interpersonal action coordination across the lifespan both from a behavioral and a neuro-cognitive perspective.

As explained above, I developed a dyadic drumming paradigm to assess interpersonal action synchronization. This part is divided into four sections: First, I will give a description of the recruitment and sample characteristics. Afterwards, I will explain the general procedure of the empirical investigation. The third section provides details on the measures used in the present dissertation and on the technical experimental setup. Finally, I will present an outline of the strategy used to statistically analyze the data.

2.1 Recruitment and Participants

The study sample consisted of 72 female participants from four different age groups: younger and older children, and younger and older adults (N = 18 per age group). I recruited participants from the participant pool of the Center for Lifespan Psychology (MPI for Human Development, Berlin), newspaper advertisements, and through posters and flyers distributed in kindergartens and sports clubs in Berlin. Children included were required to be already or become 5 years old (or 12, respectively), but not 6 years old (13 years) before the end of participation in the study. Younger adults' ages ranged from 20 to 29 years; older adults were 68–78 years old.[3] For two reasons, I only included female participants into the sample: to control for (a) gender-related differences in individuals' synchronization abilities in childhood (e.g., Hiscock et al., 1985; Wolff & Hurwitz, 1976) and (b) expected differences in interaction processes related to gender (in same and mixed-sex dyads). Additional requirements screened for in telephone interviews prior to participation were that all participants were (a) right-handed, (b) without active musical experience (i.e., no learned instrument for children and at least no practice in the last ten years for adults), (c) had good hearing, and (d) full functional mobility in both hands.[4]

[3] One older woman named a wrong year of birth in the telephone interview. Though she turned out to be younger than 70 years after the end of the study, she was still included in the older-adult age group.
[4] A total of four participants were replaced after the baseline sessions: three women from the oldest age group could no longer participate due to severe health problems and the mother of one younger girl could not bring her daughter to the laboratory regularly.

Participants (or their parents) were informed about the monetary compensation that they would receive at the end of the study (younger children: 60.00 €; older children & adults: 75.00 €). In addition to money, younger children got a sticker book to which they could add stickers after each session.

Table 2.1
Socio-Demographic Characteristics of the Sample by Age Group

	Younger Children *n* = 18		Older Children *n* = 18		Younger Adults *n* = 18		Older Adults *n* = 18	
Age (in years)								
Range	4.9–5.9		12.2–12.9		20.2–28.7		67.8–78.4	
M	5.3		12.5		25.3		73.6	
SD	0.3		0.2		2.6		2.9	
Marital Status								
Single	–		–		7	(39%)	3	(17%)
Married, living together	–		–		6	(33%)	7	(39%)
Partnership	–		–		5	(28%)	–	
Divorced	–		–		–		5	(28%)
Widowed	–		–		–		3	(17%)
Educational Level								
Kindergarten	2	(11%)	–		–		–	
Primary Education[a]	16	(89%)	–		–		1	(6%)
Lower secondary education[b]	–		6	(33%)	–		6	(33%)
High School (12th/ 13th grade)[c]	–		12	(67%)	12	(67%)	3	(17%)
Technical college/University[d]	–		–		6	(33%)	6	(33%)
Others	–		–		–		2	(11%)
Current Occupation								
Full-time employed	–		–		2	(11%)	–	
Part-time employed	–		–		2	(11%)	–	
Unemployed	–		–		1	(6%)	–	
University Student	–		–		13	(72%)	–	
Retired	–		–		–		18	(100%)

Notes. (a) German: Volks-/Hauptschule, (b) German: Mittlere Reife/Realschule, (c) German: (Fach-)Abitur, (d) German: Universität

Table 2.1 gives an overview of the socio-demographic characteristics of the sample. Educational level for children describes their current educational status, whereas the adults' highest educational degree is listed. Most of the younger children (89%) had just entered elementary school and all older children were currently in sixth grade, which means that they went to lower (33%) or higher (67%) secondary educational schools. The majority of the younger adults were university students (72%), whereas all older adults were retired. Overall, younger adults had higher educational status than older adults: They were all high school graduates or graduates of a higher educational program, whereas only 50% of the older adults held a comparable degree.

2.2 General Procedure

Except for younger children, each participant took part in seven weekly sessions[5] (see Figure 2.1). Participants started with a baseline assessment (Baseline I) where socio-demographic information, self and personality questionnaires, and cognitive tasks were assessed in small groups of two or three participants. The first baseline session was followed by an individual drumming session (Baseline II). Here, participants were introduced to the drumming paradigm and, among other tasks, individual synchronization abilities with different metronome frequencies were assessed. This session was repeated after the four dyadic sessions in which participants drummed together with one partner of each age group in randomized order (i.e., each individual was paired in one same-age and three age-mixed dyads; N = 144 dyadic sessions).

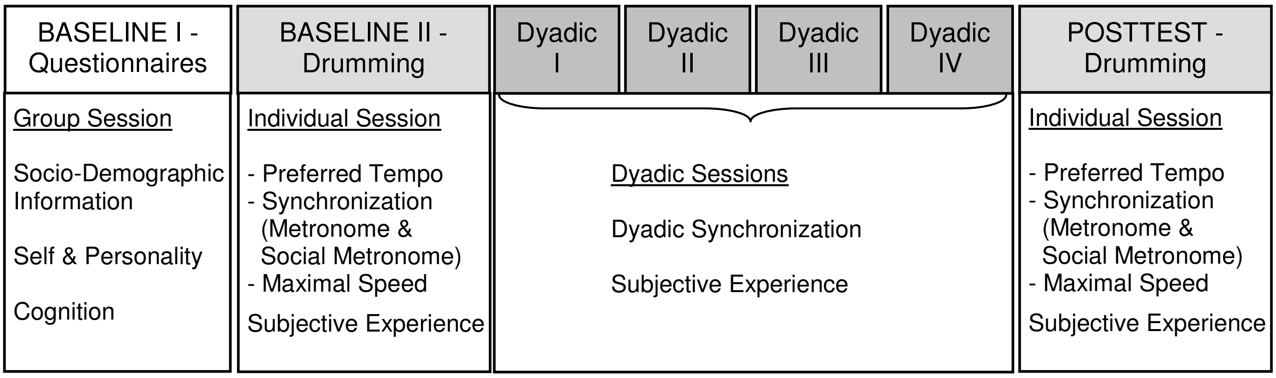

Figure 2.1. Overview of the general design.

I ensured that partners did not know each other before the beginning of the study and did not accidentally meet when joining a group session or participating in another dyadic session. In addition, by applying a complex assignment procedure, I made sure that two partners of a dyad did not share other partners (see Appendix 6.1.1, Table A1). I will provide a more specific description of the measures used for the dissertation study in a subsequent section.

Given the availability of two laboratories with drumming equipment, there were two time-shifted waves of data collection (Wave 1: 12 participants per age group; Wave 2: 6 participants per age group) adding up to a total study duration of four months. Mean time between sessions was eight days (see Table 2.2), while the mean interval between Baseline II and the posttest was about 40 days. The date was dependent on participants' (and their parents') availability. However, the range in time between sessions did not differ significantly by age-group composition, $F(9, 288) = .39$, *n.s.*

[5] Five-year-olds did not take part in the baseline questionnaire sessions and therefore only participated in 6 sessions.

Due to technical difficulties, I had to repeat seven dyadic sessions. Therefore 14 participants were involved in an 8th session. Repetition of sessions was unrelated to age group, $\chi^2 = .36$, *df* = 3, *n.s.*, and dyadic age-group composition, $\chi^2 = 7.06$, *df* = 9, *n.s.*

Table 2.2
Mean Time Between Sessions in Days

Time between Sessions:	**Mean (*SD*)**
Baseline II to Dyadic I	7.04 (2.20)
Dyadic I to Dyadic II	8.59 (4.53)
Dyadic II to Dyadic III	9.73 (5.67)
Dyadic III to Dyadic IV	9.53 (5.55)
Dyadic IV to Posttest-Drumming	7.32 (6.04)
Baseline II to Posttest-Drumming	40.02 (6.47)
Mean across all sessions	8.36 (5.07)

Baseline-group sessions with older children or adults were scheduled for two hours, whereas individual and dyadic drumming sessions ended after approximately 60 minutes. Questionnaire measures were filled out by adult participants independently. For children, however, questions and scales were read out loud and answered together. All experimenters were female student assistants who were trained to use standardized instructions in all sessions.

2.3 Measures

In the following section, I will provide details on the measures. The section will be divided into two parts: First, I will focus on central individual covariates assessed through questionnaires (e.g., self-report and others' report measures). The second section provides information on the experimental investigation of individual as well as dyadic synchronization performance including the operationalizations of individual and dyadic asynchrony.

2.3.1 Questionnaire Measures on Social Competencies and Attitudes

In the following section, I will give an overview of the self- and others' report measures that were included in the analyses of this dissertation as central indicators of the pragmatic component in the development of interpersonal action synchronization (see Table 2.3). First, I will present self-report scales that were assessed during the two baseline sessions. Others' report questionnaires which were sent to relatives named by participants and returned per mail in prepaid envelopes will be introduced afterwards. As shown in Table 2.3, not all measures were obtained for all age groups. The fact that some questionnaires were only filled out by adults or children, respectively, will be further discussed in a later section. [6]

Self-Report Scales on Situational Flexibility and Age-Specific Stereotypes

Situational flexibility. The Battery of Interpersonal Capabilities (BIC) is a self-report scale to assess functional capabilities in the interpersonal domain (Paulhus & Martin, 1987). The scale has also been used to measure individual differences in *situational flexibility* in different interpersonal contexts (Paulhus & Martin, 1988).

Adult participants were asked four questions about their capability of performing each of 16 interpersonal behaviors (e.g., dominant, trusting, etc.) in appropriate situations. One item, namely *flexible*, was added to the scale. A list of all behaviors of interest can be found in Appendix 6.1.2, Table A3. For each attribute, participants were asked a direct capability question, for example, "How capable are you of being dominant when the situation requires it?." Answers on all attributes were given on a 5-point scale ranging from 1 *(not at all)* to 5 *(very much)*. Three additional questions asked about (a) the difficulty performing each behavior, (b) anxiety when performing each behavior, and (c) the tendency to avoid situations demanding such behavior.

For each question, scores across the 17 items were aggregated to reach four composites: (1) capability, (2) difficulty, (3) anxiety, and (4) avoidance composite. As suggested by Paulhus and Martin (1988) the *capability composite* of the BIC was used as a situational flexibility index for the present dissertation. The internal consistency of this subscale was Cronbach's α = .83 (M = 38.8, SD = 9.1).

Age-specific stereotypes (AGED). I developed a new scale to assess global age-specific stereotypic expectations. This scale was based on two existing scales that both refer to *attitudes towards the aged* (Knox, Gekoski, & Kelly, 1995; Netz & Ben-Sira, 1993). In the two former scales, participants rated "typical" or "average" target persons of a specific age-group dependent on lists of bipolar adjectives (e.g., "active – passive", "flexible – inflexible"). I reassembled the adjective-lists of both scales with respect to the interpersonal context I was mainly interested in. A 5-point scale was used to rate a total of 26 pairs of bipolar adjectives. (For a complete list of all adjectives used, see Appendix 6.1.2, Table A3.) Each set of items was rated for a "typical 5-year-old", a

[6] A complete list of all questionnaire measures assessed can be found in Appendix 6.1.2, Table A2. (Note that not all of them were taken into account in the analyses).

Table 2.3
Overview of Questionnaire Measures on Social Competencies and Attitudes per Age Group

Age Group	Construct	Instrument	*n* Items	Authors/Source
5-year-olds		Others' Report		
	Parents: Interpersonal Flexibility	Interpersonal Flexibility Scale	19	Newly developed
	Kindergarten nurse: Social Skills	Social Skills Rating System (SSRS)	28	Gresham & Elliott (1990)
12-year-olds		Self-Report		
	Age-Specific Stereotypes	AGED	104	Adapted from: Knox et al., (1995); Netz & Ben-Sira, (1993)
		Others' Report		
	Parents: Interpersonal Flexibility	Interpersonal Flexibility Scale	19	Newly developed
	Teachers: Social Skills	Social Skills Rating System (SSRS)	30	Gresham & Elliott (1990)
Adults		Self-Report		
	Situational Flexibility	Battery of Interpersonal Capabilities	17	Paulhus & Martin (1987, 1988)
	Age-Specific Stereotypes	AGED	104	Adapted from: Knox et al., (1995); Netz & Ben-Sira, (1993)
		Others' Report		
	Relatives: Interpersonal Flexibility	Interpersonal Flexibility Scale	19	Newly developed

"typical 12-year-old," a "typical person in his/her 20s," and a "typical person in his/her 70s," respectively. All participants except 5-year-olds filled out this scale. Mean aggregates were computed for each target age group to obtain the following composites: *AGED-5* (Cronbach's α = .79, *M* = 2.69, *SD* = 0.39), *AGED*-12 (Cronbach's α = .88, *M* = 2.55, *SD* = 0.44), and *AGED-20* (Cronbach's α = .84, *M* = 2.50, *SD* = 0.40), and *AGED-70* (Cronbach's α = .86, *M* = 2.72, *SD* = 0.44). Higher values indicated more positive associated attitudes. Due to the small sample size, it was not possible to analyze a factor structure of the scale. Theoretical as well as methodological discussions of underlying factors in the two original scales can be found in Knox et al. (1995) and Netz and Ben-Sira (1993).

Others' Report on Interpersonal Flexibility and Social Skills

Interpersonal flexibility. To gain information on interpersonal flexibility for participants of all age groups, I developed a new scale. Because I could not assess this construct for the children using self-report measures, I developed a questionnaire to be completed by relatives of all participants. The literature includes a couple of self-report scales addressing this construct (see above), but none of them (a) included interaction processes between different age groups or (b) temporal aspects of interactions. Both aspects were very important for the present study. In 19 items, relatives of the participants were asked to rate the competency to adjust to individuals of different ages in different situations. Examples of items are: *"She can easily adapt when playing with a child"*; *"When going for a walk with an elderly person, she can easily adjust her tempo"* (the original wording of all items can be found in Appendix 6.1.2., Table A3). Answers on all items were given on a 5-point scale ranging from 1 *(not at all)* to 5 *(very much).* This scale theoretically comprised different subscales: It asked for the ability of adaptation in three contexts (i.e., play, walking, and conversation) with regard to three age groups (i.e., child, younger person, and elderly person). Again, it was not possible to analyze underlying factor structures statistically because of the small sample size. Items were averaged to obtain a mean score for *interpersonal flexibility* per participant. Internal consistency for the scale was Cronbach's α = .87 (*M* = 3.97, *SD* = 0.41).

Parents filled out questionnaires with regard to their children. Adult participants were asked to name two relatives or close friends to whom I sent the questionnaires by mail. They received monetary compensation (5.00 €) for filling out the questionnaires and sent them back in prepaid envelopes. The returned questionnaires were filled out by partners (19%), children (29%), other family members (11%), and close friends or others (non-family; 41%). The total return rate was 98%.

Social Skills Rating System (SSRS). This rating system assesses social skills of children as observed by teachers or kindergarten nurses (Gresham & Elliott, 1990). Questions focus on children's competence, assertion, and self-control when acting alone or interacting with others. Examples are: *"The child makes compromises in order to reach agreements"*; *"The child offers help to same-age children"* (for a complete list of the items see Appendix 6.1.2., Table A3). Each of the 30 questions were rated on two 3-point scales: *Frequency* (1 = never; 3 = very often) and *importance* (1 = irrelevant; 3 = fundamental) of observed behavior. For the present study, I used a *frequency composite* as an aggregate across all ratings on the frequency domain as a measure of children's social skills, Cronbach's $\alpha = .80$ ($M = 2.49$, $SD = 0.21$; 28 Items).

Questionnaires were sent out to teachers (for the 12-year-olds; $n = 15$) and to kindergarten nurses (for the 5-year-olds; $n = 18$), who received a monetary compensation (5.00 €) and were asked to send the completed questionnaires back in prepaid envelopes. The return rate was 100%.

2.3.2 Central Measures in Individual and Dyadic Drumming Sessions

In the following sections, I will introduce the measures obtained in the drumming sessions. For this, it is first necessary to give a detailed description of the technical implementation of the drumming paradigm. Second, I will describe the different conditions participants had to go through in the individual and dyadic drumming sessions. Third, the questionnaires on the subjective experience of the drumming partner and the situation will be introduced.

Technical Implementation of the Dyadic Drumming Paradigm

As described in detail in Section 1.4, I implemented an adapted version of the widely used tapping paradigm (e.g., Aschersleben, 1994, 2002; Repp, 2005) to investigate the development of interpersonal action synchronization in a controlled way. As mentioned above, the dyadic drumming paradigm had several advantages. In individual and dyadic conditions, participants were instructed to drum with a drumstick on electronic drum pads. As compared to finger tapping, drum movements require less fine motor skills, especially when being carried out with the whole forearm. Thus, this paradigm can be applied to examine synchronization abilities across different age groups, as it controls for age differences in fine motor skills. In addition, this implementation had the advantage that electronic drum pads only produced digitalized auditory drum beats into soundproof headphones while minimizing feedback from the manual drum beat itself. Therefore, the loudness of drum sounds was well controlled across participants and each beat was distinguishable from the other person's.

Figure 2.2 presents the technical set-up of the hardware used in the drumming equipment. To collect the drumming data, I used a personal computer (PC) with an Intel Pentium® 4 processor (2.8 GHz; 1 GB RAM, Windows XP Service Pack 2). Acceleration sensors, Sen1 and Sen2 (BIOVISION, single axis, sensitivity: 50 g), attached to the top end of the drumsticks were used to measure their movements. Data were recorded with a data logging card (National Instruments® M 16 E; Range: -10/+10 V).

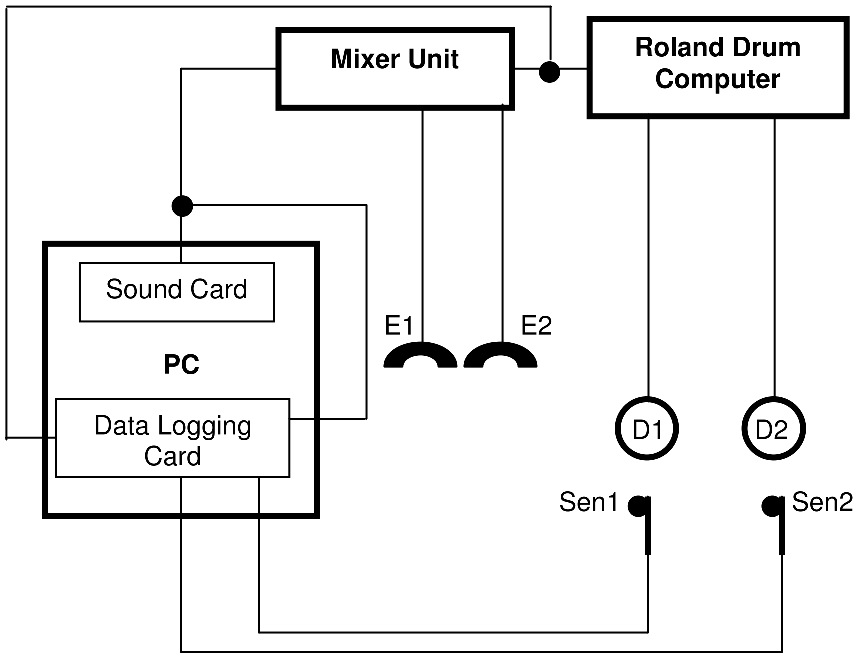

Figure 2.2. Model depicting the technical set-up of the drumming equipment (E1 = Earphone 1; E2 = Earphone 2; D1 = Drum pad 1; D2 = Drum pad 2; Sen1 = Acceleration sensor 1; Sen2 = Acceleration sensor 2).

Independent of the PC, drum sounds were produced by a drum computer (Roland® Percussion Sound Module TD-3) and presented via earphones (Numark® PHX Professional DJ-Headphones; E1 and E2), which were highly soundproof to avoid distractions through any other sound. The sensitivity of the drum pads (D1, D2) was kept at the highest value to produce the same sound intensity even when a participant drummed softly. This allowed control for volume of the drum sound, which was especially important in the dyadic session, when participants could vary in their strength of drumming. In conditions where participants heard a computer-generated drum sound to synchronize with, sounds (recorded from the drum computer) were produced by a soundcard (SoundMax Digital Audio) and were transferred through a mixer unit into the earphones.[7]

Individual Drumming Sessions

The goal of the individual drumming session was to investigate individuals' ability to drum with a mechanical time keeper, that is, with computer-generated metronome-like stimuli. The first individual session therefore started with a detailed explanation of the drumming equipment, which was separated from the experimenter by a dividing wall.[8] Drum pads and stools were personally adjusted so that participants could comfortably reach the drum with their legs positioned at a right angle at the knees. Participants were explicitly shown that drumming on the drum pads did not produce much noise, but that drum sounds could be heard through the headphones at a constant intensity even when drumming softly. The sound of participants' drum beats was a low sound (S1; *"Conga Closed Slap"*). I also provided an explanation of the use of the acceleration sensors. Participants were instructed to move the whole forearm and to drum at medium strength. All trials started with a computer-generated start signal (1200 Hz Sinus, 200 ms) and ended with an end signal (400 Hz Sinus, 200 ms).

In the following section, the different individual drumming conditions will be described. (For exact wording of the instructions in German and additional conditions not used for the current study, see Appendix 6.1.3.)

Preferred tempo. For the first condition, a record of a children's song was presented to all participants. The purpose of this song was mainly to explain to children (a) the concept of a constant meter and (b) the opportunity to use different tempi. After the song, participants were asked to decide to drum at a stable frequency that they felt most comfortable with. This task was repeated four times. Each trial ended after at least 60 drum beats <u>and</u> 45 seconds of drumming were recorded (i.e., both conditions needed to be met; for the score and text of the song, see Appendix 6.1.3).[9]

Synchronization with a metronome. In the individual synchronization conditions, computer-generated stable metronome-like signals were presented through the headphones at a sound of a

[7] The standard deviation of the intervals between metronome signals due to equipment imprecision was less than 2 ms.
[8] I used cameras to control the behavior of the participants online.
[9] Mean inter-drum intervals across all sessions ranged from 276 to 1283 ms (M = 545; for distribution and age-group differences, see Appendix 6.1.4). Inconsistent with the literature, significant differences in the mean preferred tempo between sessions were found in a repeated measures analysis of variance with within-subjects factor measurement occasions (4), $F(3, 71) = 14.51$, $p < .01$, $\eta^2 = .17$. This indicates that in the present sample individuals' preferred tempi were not stable across time.

higher drum (S2; *"Wood Block Hi"*), which was easily distinguishable from the participants' own lower drum sound (S1). The instruction was to listen to the computer-generated drum sound and to drum in a synchronized manner with it as soon as possible. A single trial at an inter-stimulus interval (ISI) of 500 ms (duration: 45 s; 90 signals) allowed first practice. After this, the synchronization task was repeated with ISIs of 419 ms, 757 ms, and the value of the participant's mean preferred tempo (operationalized as the mean inter-drum interval (IDI) in the preferred tempo condition). The values of the two fixed ISIs were chosen as prime numbers to make sure that they were not interrelated, that is, not divisible by the same value. Each block consisted of four trials. The criterion for trial length was at least 60 computer-generated drum beats and 45 s of drumming recorded. The order of blocks in the synchronization conditions was kept constant across participants.

Synchronization with a social metronome. The social metronome condition was basically very similar to the synchronization condition with a metronome. In this case, participants were told that they would be presented the recordings of other individuals drumming at their respective preferred tempo. Again, they were asked to listen to the recorded drum sound of the other person and to drum in synchrony with it as soon as possible. In this condition, the ISIs of the presented drum sounds were not stable, but showed human-like variability around a specified mean ISI. The algorithm to produce this variability was based on results by Krampe et al. (2005), who reported a positive correlation between variability of inter-tap intervals in a continued tapping task and tapping tempo. From the same study, I used findings on age group differences in continued tapping stability to assign two degrees of variability: one associated with younger adults (i.e., lower variability) and one associated with older adults (i.e., higher variability). Every ISI also used in the synchronization condition (419 ms, 757 ms, and mean preferred tempo) was presented at each degree of variability.[10] The order of the resulting six blocks (four trials each with at least 60 signals and 45 s recorded drumming) of social metronome conditions was assigned randomly across participants.

Maximal speed. In the last condition in the individual sessions, participants were instructed to drum as fast as possible for 15 s. This task was again repeated four times.

[10] The ISIs presented were randomly and independently drawn from normal distributions with M = ISI and variances $ISI_{younger} = 0.00163 \cdot ISI - 243$ and $ISI_{older} = 0.00134 \cdot ISI - 99$. Each ISI was only based on the previous stimulus, that is, the correlation between the lags was zero.

Dyadic Sessions

All dyadic sessions began with the assessment of each participant's *preferred tempo* (see above) while the drumming partner waited outside the room. Afterwards, both participants were briefly introduced to each other before being seated in front of one of two drum pads which were separated with dividing walls (see Figure 2.3). Each participant drummed shortly on her drum pad to try out the drum sounds: one participant got a low (S1) and the other participant a high drum sound (S2). The low drum sound was the same sound that participants had already used in the individual sessions, whereas the high drum sound was the sound which was formerly used as the metronome signal. Each participant was assigned to the low and high drum sound twice respectively in the four dyadic sessions. The order of assignment was randomized across participants. The dyadic drumming condition comprised two blocks with eight trials each in which participants were instructed to drum in synchrony with each other at a constant frequency that they agreed upon without talking to each other. Again, the trial duration was at least 45 s of recorded drumming <u>and</u> 60 signals from at least one participant.

Figure 2.3. Example of drumming session setting.

Self-Report Measures in Individual and Dyadic Drumming Sessions

To assess differences in the subjective experience of the drumming situation as well as the interaction partner, participants were asked to fill out several questionnaires before, during, or after the session (see Table 2.4). The following ratings were used to operationalize individuals' experiences as affected by interpersonal action synchronization:

Table 2.4
Drumming Sessions Protocol (Sessions lasted approximately one hour)

Session	**Conditions included**
Individual Sessions – Baseline II & Posttest	Internal Tempo (4 Trials)
	Synchronization with Metronome (3 Blocks, 4 Trials)
	Subjective Rating of Synchronization Accuracy After each Block*
	Synchronization with Social Metronome (6 Blocks, 4 Trials)
	Subjective Rating of Synchronization Accuracy After each Block*
	Maximal Speed (4 Trials)
	Subjective Rating of Synchronization Accuracy of Session / Feedback*
	Feedback
Dyadic Sessions	Internal Tempo (4 Trials)
	First Impression*
	Dyadic Synchronization Block 1 (8 Trials)
	Subjective Rating of Dyadic Synchronization Accuracy: Block 1*
	Dyadic Synchronization Block 2 (8 Trials)
	Subjective Rating of Dyadic Synchronization Accuracy: Block 2 & Session*
	Last Impression Questionnaire*
	Feedback

Note. Variables/tasks central to this dissertation are highlighted in bold. * Questionnaire measures.

First & Last Impression. In the dyadic sessions, after being introduced to each other, participants rated their first impression of each drumming partner in six items (see Appendix 6.1.5, Table A5) on a 5-point scale *(very much–not at all)*. Based on theoretical considerations, the four questions of positive partner evaluation *("What do you think – how likeable/friendly/cooperative is today's drumming partner?"; "How much would you like to get to know today's drumming partner better?")* were aggregated as a measure of *First Impression*, Cronbach's α = .73 (M = 2.31, SD = 0.70).

At the end of the session, a similar questionnaire was presented again, extended by ten additional questions (see Appendix 6.1.5, Table A5). A *Last Impression* aggregate was calculated across the equivalent four items *"How likeable/friendly/cooperative was today's drumming partner?"* and *"How much would you like to get to know today's drumming partner better?"*, Cronbach's α = .78 (M = 2.27, SD = 0.75). In additional single items, participants were also asked about their satisfaction with the drumming performance, their perceived difficulty in drumming with the respective partner, as well as how positively they would rate the drumming situation. For a complete list of items that were not further analyzed in the present study see Appendix 6.1.5, Table A4. All items in the final questionnaire were rated using a 5-point scale *(very much–not at all)*.

Rating of own synchronization accuracy. After each block and at the end of each session, participants rated their perception of synchronization accuracy answering the question *"How well did you succeed in synchronizing with the other drum/person?"* on a 10-tier bar which had to be colored manually (see Section 6.1.6): The higher the perceived accuracy of a participant, the more tiers were colored. This scale was used to help younger children rate their synchronization accuracy.[11]

Enjoyment during synchronization performance. A single-item measure was used *("How much did you enjoy drumming with the other drum/person?")* to examine positive affect during the drumming tasks. Answers on this item were given on a 5-point scale that ranged from 1 *(very much)* to 5 *(not at all)* by all age groups except 5-year-olds.

Feedback. At the end of each session, experimenters provided blockwise feedback of synchronization accuracy during the respective conditions. Feedback was presented on the computer screen on colored bars like those used for the subjective rating.

2.3.3 Operationalization of Asynchrony

In this section, I will explain the newly developed measure to operationalize *individual* and *dyadic asynchrony*. In general, the maxima of the acceleration distribution recorded by the sensors were defined as drum beats. The synchronization accuracy when drumming with a metronome or another person, respectively, was operationalized by a *measure of asynchrony* (von Oertzen, 2006). The measure of asynchrony compares two time-series of drum beats, one from person A and one from the metronome or person B, respectively. The underlying idea was to calculate the distance between the two series as costs of transforming one series into the other, that is, to reach perfect synchrony.

To this end, an algorithm was developed that transfers one drumming sequence to the other by either shifting drum beats to later or earlier points in time, or by inserting or deleting drum beats. The algorithm calculates the costs for a time shift in milliseconds and the costs for insertion or deletion of drum beats as half the mean inter-drum-beat interval of the series. It automatically pairs drum beats such that an optimal trade-off between shifting and inserting missing drum beats is assumed, that is, the algorithm finds the transformation with minimal costs. This optimization is achieved by dynamic programming (e.g., Corman, Leierson, & Rivest, 1994).

[11] After the end of the study, rating data of 5-year-olds were excluded from analyses because observation during test sessions indicated that the understanding of the rating scales was poor in this age group.

An example of the procedure is shown in Figure 2.4: Each drum beat of Series 2 is moved in time towards a drum beat of Series 1. A "penalty cost" for deleting/inserting drum beats is exemplarily set to 200 ms.[12] The underlying algorithm continuously compares the costs for moving (in ms) with this fixed penalty cost and then executes the transformation with lower costs. For example, o_1 is moved in time to the position of x_1. Assigning o_2 to x_2 would cause higher costs (approximately 450 ms) than deleting o_2 (200 ms) and inserting a new drum beat (o_{2new}; 200 ms) at the x_2 point in time. The sum of all costs (i.e., 810 in the example) is used as a *measure of asynchrony*. (A more detailed description of the algorithm can be found in Appendix 6.1.7., Figure A2.)

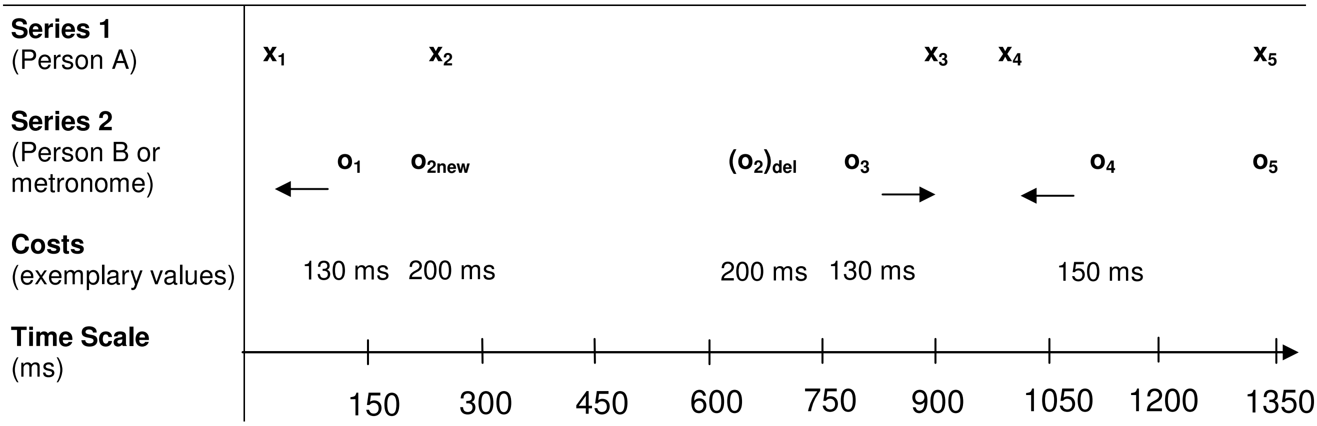

Figure 2.4. Example of procedure to calculate the measure of asynchrony.

The scale of this measure is symmetrical (i.e., it is unimportant which time series functions as Series 1 or Series 2) and can be applied to relatively small time windows because it automatically produces asynchrony values for each point in time. It is therefore possible to use detailed information on the temporal dynamics of the process at high time resolution. In this dissertation, however, analyses are restricted to mean levels, that is, asynchrony across within-trial sums for specific conditions. The resulting measures are referred to as *asynchrony with metronome* (i.e., mean across all metronome conditions in the baseline session), *asynchrony with social metronome* (i.e., mean across all social-metronome conditions in the baseline session), and *dyadic asynchrony* (i.e., mean across all trials in a dyadic session).[13] Because the two individual measures of asynchrony (i.e., asynchrony with metronome and social metronome) were highly correlated

[12] In effect, penalty costs are calculated individually within each trial, among others, depending on the tempo (i.e., as half of the mean IDI of the participants).

[13] There is no individual measure of asynchrony in the dyadic condition, that is, both partners receive the same asynchrony value.

($r = .93$, $p < .01$), a mean aggregate of *individual asynchrony* was calculated as an additional individual measure of asynchrony. This aggregated measure was implemented as an operationalization of individuals' synchronization abilities (i.e., an indicator of sensorimotor competencies). Dyadic asynchrony was used to operationalize the dyadic performance (i.e., interpersonal action synchronization accuracy).[14]

2.4 General Statistical Procedures

In the following, I will introduce the general statistical procedures that were used to analyze the data. First, I will provide an overview of the specific data structure in the present study. Afterwards, I will explain how multilevel modeling with Bayesian estimation methods was applied to analyze the first two sets of hypotheses. As this statistical approach is still rather new, I will describe it in more detail in the subsequent section. Third, I will introduce the functional equations of the models that were tested. Finally, I will address the distributions of the variables that were included in the analyses, the implemented centering procedures, and the treatment of missing values.

All statistical analyses were conducted using SPSS 15.0 for Windows (SPPS Inc., 2006), R 2.4.1 for Windows (R Development Core Team, 2007), WinBUGS 1.4.1 (Lunn, Thomas, Best, & Spiegelhalter, 2000), and SAS 9.1 for Windows (SAS Institute Inc., 2003). The software used for the respective analyses will always be indicated when presenting the results.

2.4.1 Data Structure

The data of the present study had a rather special structure, as each individual was included in four different dyads (one same-age and three age-mixed dyads). To analyze the data, it was necessary to relate variance in the dyadic performance to (a) differences between dyads and (b) differences between individuals in the dyads. Hence, multilevel modeling techniques were used to capture the hierarchical structure in the data (Hox, 2002; Snijders & Bosker, 1999). For example, the multilevel models used for research questions I and II considered two levels of analyses: the dyadic level (Level 1) and the individual level (Level 2), which further specified differences within the dyads.

[14] A repeated measures analysis of variance with the within-subjects variable measurement occasions (4) was conducted to control for training effects. Results indicated that there was no significant change (e.g., improvement) in mean level of interpersonal action synchronization across the four measurement occasions, $F(3, 71) = 0.135$, *n.s.*

A further characteristic of the data structure was the dependency between dyads: As each person was included in four dyads, a *multiple membership structure* had to be taken into consideration when setting up a model (e.g., Browne, 2005; Fielding & Goldstein, 2006). In medical studies, for example, a hospital patient may be treated by several nurses and each nurse may then have an effect on the patient's recovery. Analyses taking into account multiple membership structures, can answer questions on the impact of different nurses to explain the final recovery progress (e.g., Browne, Goldstein, & Rasbash, 2001). In the case of the present study, the dyadic outcome (i.e., the dyadic asynchrony in one dyad) was dependent on different individuals and each individual was included in four dyads (see Figure 2.5).

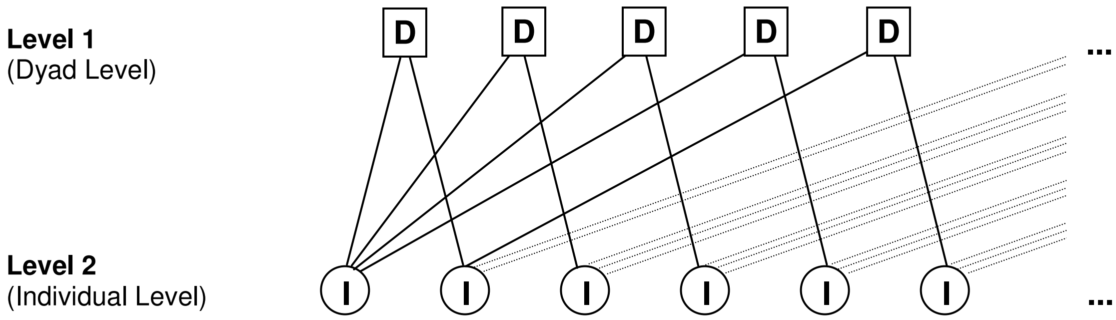

Figure 2.5. *Schematic illustration of levels of analyses: Each individual included in four dyads.*

To summarize, analysis of the data with regard to the first two sets of research questions (i.e., prediction of differences in dyadic synchronization accuracy), required specification of models (a) that identify variance components explained by differences between dyads and individuals, respectively, (b) in which further characteristics of both dyads and individuals can be included, and (c) that take into account the dependencies between dyads. It was possible to set up these highly complex models in R (R Development Core Team, 2007) and estimate them using an estimation method based on Bayesian inference statistics in WinBUGS (**Win**dows **B**ayesian Inference **U**sing **G**ibbs **S**ampling; Lunn et al., 2000). The strength of Bayesian parameter estimation procedures lies in their ability to deal with small sample sizes and complex statistical models with a high number of parameters to be estimated. If there is not enough information available to create accurate point estimates of variance parameters, information on this inaccuracy or estimation error can be obtained using a Bayesian approach (e.g., Gelman & Hill, 2007). The third research question, focusing on the relationship between dyadic synchronization accuracy and individuals' subjective experience within a session, was analyzed within a repeated measures design in SAS 9.1 for Windows (SAS Institute Inc., 2003) based on general frequency statistics. These models will be further described below.

2.4.2 Bayesian Parameter Estimation Methods

In the literature on multilevel modeling, two main approaches of statistical models are discussed. Most often, the estimation of parameters in multilevel modeling is carried out by likelihood-based frequency statistics, where population-based inferences are made from sample data based on a maximum likelihood (ML) algorithm. As an alternative, Bayesian estimation procedures are able to include prior knowledge about the question of interest (e.g., from previous research) into the estimation process and can therefore be used to fit many models in which ML procedures could lead to non-reliable results. As Bayesian estimation procedures are still rather new, I will give a basic overview of the assumptions and underlying routines in the next section. I will also introduce the estimation criteria that will be used later to provide information on model fit and parameter estimation accuracy.

The Bayesian approach combines prior beliefs about the parameters with the data collected to produce new posterior beliefs. To begin with, there is always a set of unknown parameters (Θ) for which prior beliefs can be condensed into prior distributions, $p(\Theta)$. Prior distributions reflect previous knowledge about the parameters. Based on the observed data y a likelihood function $L(y|\Theta)$ can be determined (i.e., in analogy to the function that is maximized in maximum likelihood estimation). If a "non-informative" prior distribution is selected (i.e., there is only marginal previous knowledge about the parameter of interest), the likelihood function and the posterior distribution will essentially be the same. The prior and the posterior distributions are combined to form a joint posterior distribution for Θ, that is $p(\Theta|y) \sim p(\Theta)L(y|\Theta)$. Inferences about Θ are finally reached from this distribution (Browne, 2005).

The Bayesian estimation method used in WinBUGS is a *Markov Chain Monte Carlo* (MCMC) estimation. Instead of using a relatively complex joint posterior distribution, MCMC methods generate a large number of simulated random draws from conditional distributions of all the parameters, for example, by means of a *Gibbs Sampling* algorithm (e.g., Gill, 2002; Gelman & Hill, 2007; Spiegelhalter, Thomas, Best, & Lunn, 2003). Parameter estimations are continuously updated by drawing values from the respective distributions assuming that the current estimated values for the other parameters are the true values. The single draws from the conditional distribution can be regarded as realizations of the posterior distribution if the simulations of a conditional distribution converge to the true stationary distribution (e.g., Cowles & Carlin, 1996).

The basic principle is that once the chain has run long enough, it will approach the desired posterior distribution (Gill, 2002). It is then possible to calculate the posterior mean and standard deviation from the random draws of the parameter of interest. That is, the mean of this distribution is used as the best point estimate for the parameter. In analogue to confidence intervals in the frequentist approach, the *Bayesian Credible Interval* (BCI) is based on the 2.5th and 97.5th percentile points of the posterior distribution. With a probability of .95, the true value of the estimated parameter lies within this interval (Browne, 2005; see below). In contrast to ML statistics, the MCMC methods therefore not only provide point estimates and standard errors for all parameters, but also provide a posterior distribution of the parameters. Density plots of the complete posterior distribution supply further information on the estimated parameter. ML and MCMC approaches have now been compared in a number of studies (e.g., Browne & Draper, 2006). One important advantage of MCMC methods for the present study is that they can be generalized to fit more complex multilevel models that are not estimable using a ML approach with the software packages currently available.

As explained above, one benefit of using Bayesian estimation methods is that prior knowledge can be integrated into the model estimation, that is, prior distributions can be specified for each parameter to be estimated. However, in the current study I had no specific assumptions about the expected parameters. That is why I used *non-informative* prior distributions for all estimation procedures (i.e., by default of the program, this was set to a normal distribution with $M = 0$ and $Var = 1{,}000{,}000$).

Bayesian Estimation Criteria

There are mainly three estimation criteria that I will report in the result section to identify meaningful results of the estimation procedure: the *Bayesian Credible Interval*, the *Deviance Information Criterion*, and the *rhat* value ($\hat{R}$). These criteria will be explained in more detail in the following sections.

Bayesian Credible Interval (BCI). As a parameter estimation criterion, the BCI is the posterior probability interval in which an estimated parameter t lies with a specified probability. For example, if a 95% credible interval for a parameter t is 1.2 – 3, this means that the posterior probability that the true value of t lies in this interval is .95.[15] BCIs that do not include "0" as

[15] Note that this is a stronger statement than a *confidence interval* used in frequentist inference statistics that should not be confused with the BCI. A 95% confidence interval of 1.2 – 3 in this case would mean that with a large number of repeated samples, 95% of the calculated confidence intervals would include the true value of the parameter *t*.

possible values can be interpreted as estimated parameter values that are reliably different from zero. In the following, these will be referred to as *reliable* effects.

Deviance Information Criterion (DIC). The DIC consists of two additive components. The first component is a goodness-of-fit measure of the estimated model, that is, the mean deviance over all *n* simulated parameter vectors. The better the model fits the data, the smaller the value of this measure is. Second, it includes an additional penalty term for increasing model complexity (i.e., it specifies the effective number of parameters). Generally speaking, when comparing two or more models, a lower absolute value of the DIC can be interpreted as a model with a higher model fit. The DIC is a generalization of the Akaike Information Criterion (AIC, Akaike, 1973) and the Bayesian Information Criterion (BIC, Schwarz, 1978) for complex hierarchical models (Congdon, 2006; Gelman & Hill, 2007; Spiegelhalter, Best, Carlin, & van der Linde, 2002). It is used in Bayesian model comparison, particularly if the posterior distributions have been obtained by MCMC simulation. In models with negligible prior information, the estimation of the DIC is equivalent to the AIC. Nested as well as non-nested models can be compared in their model fit by inspecting the differences between the DIC values. However, further research is necessary to define what would constitute an important difference in DIC, especially in small samples. A preliminary rule of thumb is: $\text{Diff}_{\text{DIC1–DIC2}} \geq 10$: important difference; $\text{Diff}_{\text{DIC1–DIC2}} = 5–10$: substantial difference; $\text{Diff}_{\text{DIC1–DIC2}} < 5$: non-interpretable difference (Spiegelhalter et al., 2003).[16]

$\hat{\text{R}}$ *(rhat).* WinBUGS uses an algorithm that runs several Markov chains in parallel. Convergence is assessed by examining whether the discrepancies between the different chains decrease. For each parameter that is saved, $\hat{\text{R}}$ is, approximately, the square root of the variance of the mixture of all chains, divided by the within-chain variance. At convergence of the algorithm, $\hat{\text{R}}$ should equal 1. $\hat{\text{R}} \leq 1.1$ for all parameters is interpreted as sufficient convergence (Gelman & Hill, 2007). In addition, density distributions of the estimated chains may also provide information on the simulation process.

2.4.3 Model Sequence and Model Notation

In the following, I will provide details on the models that were used to analyze the data. For each set of research questions, a sequence of models was examined in order to compare the impact of different predictors. This sequence always included an *unconditional model* to be used as a baseline against which all models were compared. The first part of this section will focus on the

[16] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml

analyses of individual and age-related differences in dyadic asynchrony. In the second section, I will describe the analyses concerning the predictive value of dyadic asynchrony on individual outcome measures of subjective experience.

Analyzing Differences in Dyadic Asynchrony (Research Questions I & II)

The multilevel models examining the first two research questions (i.e., How do individual and age-related differences in sensorimotor abilities and social competencies relate to dyadic action synchronization? Do dyads of varying age compositions differ in dyadic action synchronization?) considered two levels of analyses: the dyadic level (Level 1) and the individual level (Level 2) that further specifies differences within the dyads. In the dyadic condition, each dyad received one specific value (i.e., dyadic asynchrony) referring to their mean asynchrony when drumming with each other. For all regression equations modeled, dyadic asynchrony was included as outcome at the dyadic level (Level 1). These models were set up in R 2.4.1 for Windows (R Development Core Team, 2007) and were estimated using WinBUGS 1.4.1 (Imperial College and MRC, 2004).

Varying-intercept model. The first analysis aimed at answering the question of how much variance in dyadic asynchrony can be explained by (a) differences between dyads and (b) differences between individuals within dyads. The simplest multilevel model separates the variance components on each level and can be formulated as follows:

$$Y_i = \beta_0 + u_j\,[\,p1_i\,] + u_j\,[\,p2_i\,] + \varepsilon_i \qquad (1)$$

with $u_j \sim N(0, \sigma^2_u)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$.

This model is a varying-intercept model with normally distributed dyadic and individual-level errors, where Y_i represents the dyadic asynchrony for the dyad i and β_0 represents the average dyadic asynchrony across all dyads in the whole sample. The model postulates the asynchrony within a dyad to be an additive effect of each individual's influence on the dyadic outcome; $p1_i$ refers to the first person in the dyad i and $p2_i$ refers to the respective second person. That is, the variability in the dyadic outcome, which is related to the differences between individuals, is divided equally between the two individuals (i.e., u_j [$p1_i$] and u_j [$p2_i$]). The average individual performance across its four dyads was extracted from the respective dyadic outcome by estimating u_j for each individual. The normal distribution of the u_j can be understood as the prior distribution for each individual's effect within the dyadic asynchrony. The mean of this distribution was $\mu_u = 0$. The respective variance σ^2_u was itself estimated from the data and is therefore called a *hyperparameter* (Gelman & Hill, 2007). The value of this variance parameter σ^2_u

was used as an estimator of the variance component between individuals. The variance component that was related to differences between dyads was indicated by the estimated value of the parameter σ^2_ε. Non-informative prior distributions were specified for all fixed and random effects in the model: $\beta_0 \sim N(0, 1{,}000{,}000)$, $\sigma^2_\varepsilon \sim \Gamma(0.001, 100)$, and $\sigma^2_u \sim \Gamma(0.001, 100)$, assuming that variances were gamma-distributed to avoid negative variances. This first model assumes variance in dyadic asynchrony to be explained solely by differences between individuals and between dyads. It is therefore possible to discriminate proportions of total variance that are related to differences between and within dyads (i.e., between individuals). Thus, this model will be used as a baseline model (i.e., an *unconditional model*) for further model comparisons.

Specifying differences at individual and dyadic level. Using multilevel modeling, it was further possible to include additional predictors at both levels: Differences in the effects of individual predictors could be specified through covariates included at Level 2 and the ten different dyadic age-group compositions were included as predictors to further explain variance between dyads at Level 1.

The second set of models accounted for different individual predictors, because the effect of individual performance on the dyadic outcome was assumed to differ due to individuals' abilities in sensorimotor skills or social competencies. In the respective multilevel models, intercepts at Level 2 were postulated to vary across individuals:

$$Y_i = \beta_0 + u_j\,[\,p1_i\,] + u_j\,[\,p2_i\,] + \varepsilon_i \qquad (2)$$

with $u_j \sim N(a \cdot V, \sigma^2_u)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$.

Several models with different individual predictors were run (e.g., individual asynchrony with metronome; others' report on interpersonal flexibility etc.). The parameter V was replaced by any variable measured at the individual level.[17] With this set of models, it was possible to estimate how different individual abilities (i.e., sensorimotor skills and social competencies) can explain variability in dyadic asynchrony. Non-informative prior distributions were specified for all fixed, $\beta_0 \sim N(0, 1{,}000{,}000)$; $a \sim N(0, 1{,}000{,}000)$, and random effects, $\sigma^2_\varepsilon \sim \Gamma(0.001, 100)$; $\sigma^2_u \sim \Gamma(0.001, 100)$.

The second research question focused on differences in dyadic asynchrony that can be explained by differences in the age-group composition of the dyads. Hence, a further model (3) included predictors at the dyadic level, that is, dummy-coded variables that referred to each of the 10 possible age-group compositions of the dyads (YC: younger child, OC: older child, YA:

[17] For two individual predictors integrated into the model, u_j was specified as: $u_j \sim N(a \cdot V_1 + b \cdot V_2, \sigma^2_u)$.

younger adult, OA: older adult; reference category: YAYA combination). In the following model, it was postulated that the intercepts varied across dyads due to the age composition of the respective dyad:

$$Y_i = \beta_0 + u_j\,[\,p1_i\,] + u_j\,[\,p2_i\,] + \beta_1 \cdot YCYC + \beta_2 \cdot YCOC \quad (3)$$
$$+ \beta_3 \cdot YCYA + \beta_4 \cdot YCOA + \beta_5 \cdot OCOC + \beta_6 \cdot OCYA$$
$$+ \beta_7 \cdot OCYA + \beta_8 \cdot OCOA + \beta_9 \cdot OAOA + \varepsilon_i$$

with $\quad u_j \sim N(0, \sigma^2_u)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$.

Again, non-informative prior distributions were specified for all fixed, $\beta_0 - \beta_9 \sim N(0, 1{,}000{,}000)$, and random effects, $\sigma^2_\varepsilon \sim \Gamma(0.001, 100)$; $\sigma^2_u \sim \Gamma(0.001, 100)$, in the model.

In Hypothesis 2c, I expected the relationship between dyadic age-group compositions and dyadic asynchrony to be predicted, in part, by differences between individuals. Due to the specific hierarchical data structure (i.e., dyadic asynchrony differed at the dyadic outcome level and individual predictors differed at the individual level) it was not possible to analyze this mediation in the way it has been introduced in the literature on multilevel mediation (e.g., Krull & McKinnon, 1999, 2001; MacKinnon, Fairchild, & Fritz, 2007). To examine the hypothesis as directly as possible, I used a final set of multilevel models (4) that included both the dyadic age-group combinations (at the dyadic level) and the respective individual predictor (at the individual level). These models were applied to analyze whether differences between age-group combinations of the dyads were predictive of dyadic asynchrony while controlling for individuals' sensorimotor skills or social competencies. Multilevel models of this type were:

$$Y_i = \beta_0 + u_j\,[\,p1_i\,] + u_j\,[\,p2_i\,] + \beta_1 \cdot YCYC + \beta_2 \cdot YCOC \quad (4)$$
$$+ \beta_3 \cdot YCYA + \beta_4 \cdot YCOA + \beta_5 \cdot OCOC + \beta_6 \cdot OCYA$$
$$+ \beta_7 \cdot OCYA + \beta_8 \cdot OCOA + \beta_9 \cdot OAOA + \varepsilon_i$$

with $\quad u_j \sim N(a \cdot V, \sigma^2_u)$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$,

where the parameter V can again be replaced by each individual predictor included. Non-informative prior distributions were specified for all fixed, $\beta_0 - \beta_9 \sim N(0, 1{,}000{,}000)$; $a \sim N(0, 1{,}000{,}000)$, and random effects, $\sigma^2_\varepsilon \sim \Gamma(0.001, 100)$; $\sigma^2_u \sim \Gamma(0.001, 100)$.

Model assumptions. Due to the small sample size and the complexity of the models, it was not directly possible to check for normal distribution and variance homogeneity of the variance components.[18] However, Gelman and Hill (2007) propose pragmatic model specifications and refer to the fact that violation of the main assumptions (i.e., normal distribution of the errors, equal variances) does not affect the most important aspects of regression models. Furthermore, as no prior information was specified in the prior distributions, parameter estimation is comparable to estimation procedures based on maximum likelihood algorithms. As general linear modeling techniques are described as relatively robust against violations of these general model assumptions, the same argumentation applies to the present models (e.g., Maxwell & Delaney, 2004; Snijders & Bosker, 1999).

Predictive Effects of Dyadic Asynchrony on Individual Outcomes (Research Question III)

The third research question asked how predictive dyadic synchronization accuracy was of the subjective experience of each individual's partner and the situation in general. In contrast to the first two research questions, related analyses included individual ratings within a session as outcome variables and were therefore modeled within a repeated-measures design with four measurement occasions per individual (four different partners from the respective age groups). These models were analyzed in SAS 9.1 for Windows (SAS Institute Inc., 2003) based on ML estimation procedures.

For each individual (except younger children), values of dyadic asynchrony and values of different measures of subjective experience (Y_{ij}) were available for each of the four measurement occasions. Therefore, the model can be formulated as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,\text{Dyadic Asynchrony}_{ij} + \varepsilon_{ij} \tag{5}$$

with $\beta_{0j} \sim N(\tau_0, \sigma^2_{\beta 0j})$, $\beta_{1j} \sim N(\tau_1, \sigma^2_{\beta 1j})$

$$\text{and } \varepsilon_{ij} \sim N(0, \Sigma),\ \sum \equiv \begin{bmatrix} \sigma^2_{\varepsilon 1} & 0 & 0 & 0 \\ 0 & \sigma^2_{\varepsilon 2} & 0 & 0 \\ 0 & 0 & \sigma^2_{\varepsilon 3} & 0 \\ 0 & 0 & 0 & \sigma^2_{\varepsilon 4} \end{bmatrix}.$$

β_{0j} represents the mean subjective experience for each individual j across all time points i. The parameter β_{1j} describes the differences in each individual's subjective experience that is related to the dyadic asynchrony within a given session. Within-person differences between measurement occasions are represented by ε_{ij}.

[18] Variance in dyadic asynchrony did not differ by age-group composition (Levene's test: F (9, 134) = 1.17, *n.s.*).

It was necessary to further control for individual differences in the subjective experience that could be related to, for example, differences in the age of the partner. Therefore, models including dummy variables at Level 1 representing the respective age group, with YA as reference category were set up. These conditional models can be described as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}\,\text{Dyadic Asynchrony}_{ij} + \beta_{2j} \cdot \text{YC}_{ij} + \beta_{3j} \cdot \text{OA}_{ij} + \beta_{4j} \cdot \text{OA}_{ij} + \varepsilon_{ij} \quad (6)$$

with $\beta_{0j} \sim N(\tau_0, \sigma^2_{\beta 0j})$, $\beta_{1j} \sim N(\tau_1, \sigma^2_{\beta 1j})$,

and $\varepsilon_{ij} \sim N(0, \Sigma)$, $\sum \equiv \begin{bmatrix} \sigma^2_{\varepsilon 1} & 0 & 0 & 0 \\ 0 & \sigma^2_{\varepsilon 2} & 0 & 0 \\ 0 & 0 & \sigma^2_{\varepsilon 3} & 0 \\ 0 & 0 & 0 & \sigma^2_{\varepsilon 4} \end{bmatrix}$.

Generally, separate models were run for each outcome variable (e.g., last impression, positive perception of the situation, satisfaction with the drumming performance, experienced difficulty when drumming). Other individual predictors of interest (e.g., first impression, interaction terms) were included similarly.

*Model assumption*s. Visual inspection of distributions of the random effects and residuals indicated that the assumption of normal distribution was approximately met. As recommended by Littell, Milliken, Stroup, Wolfinger, and Schabenberger (2006) for repeated measures-designs, the degrees of freedom were adjusted according to the Kenward-Roger (KR) correction procedure (see also Kenward & Roger, 1997). To allow for model convergence and make the models more parsimonious, the covariance between the different dyadic sessions and the respective outcome was set to "0", that is, only variances were included in the estimation procedure.

2.4.4 Variable Distributions

Using SPSS Examine, all variables were checked for significant departure from normality prior to analyses. In variables, in which significant skewness and kurtosis was detected, satisfactory approximation of normal distribution could be achieved through transformation. I followed the respective recommendations by Tabachnick and Fidell (2001). For an overview of variable distributions and information about transformations used, see Appendix 6.1.8, Table A6.

2.4.5 Centering of Predictor Variables

As recommended in the literature on multilevel modeling (e.g., Hox, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999), all continuous variables used as predictors in the present study were *grand-mean centered* (i.e., the overall mean was subtracted from all values of the variable). This allows the interpretation of the intercept as the expected value of the outcome variable when all explanatory variables have their mean value.

2.4.6 Structure and Treatment of Missing Values

It is necessary to highlight (a) the structure of missing values that occurred in the study due to the study design and (b) the particularities of treatment of missing values when using Bayesian estimation procedures (especially in WinBUGS).

Due to the wide age range of the sample, it was not possible to obtain the same covariate measures for each of the four age groups, that is, 5- and 12-year-olds were not asked to fill out the same self-report scales as younger and older adults (see Section 2.3.1). This meant that several measures were only available for half of the dataset (e.g., *Situational Flexibility Scale* only for adults, *Social Skills Rating System* only for children). This huge proportion of *planned* missing data needed to be taken into account in several analyses.

Missing values in WinBUGS. In WinBUGS, missing values must be explicitly modeled. Cases with missing data need to be excluded before models can be specified or variables with missing values must be modeled explicitly. In Bayesian statistics, every unknown parameter must have a defined prior distribution. Any missing value is treated as a parameter to be estimated and therefore needs to be assigned a priori. Therefore, distributions of missing values have to be defined because sampling requires the full conditional distributions to be specified (e.g., Gelman & Hill, 2007). This can lead to very biased models, especially when estimation in imputation models is based on small sub-samples that would theoretically show different characteristics on the variables of interest. I therefore finally decided against the imputation of missing values, because this would have entailed the estimation of values for a complete age group based on observed values from another age group. Hence, some analyses were conducted on different sub-samples without missing values on the respective predictor variable. Some follow-up analyses including covariates with high proportions of missing values therefore only referred to very small datasets. These analyses, of course, only apply to the age groups actually analyzed, and they have to be interpreted with caution, especially in view of the fact that some effects did not reach the level of statistical significance.

Missing values in SAS. The analyses that were run in SAS used dyadic asynchrony to predict different individual outcome variables on individuals' subjective experience of the partner and the interaction situation. These measures were not available for younger children, because they did not fill out final questionnaires. Therefore, analyses were only run for valid cases on the outcome variables.