Supplement 3. Power analysis.

In an age-comparative study with subjects ranging in age from 20 to 80+ years, we plan to have subjects walk on a beam (length: 4 m, height: 2cm) of varying widths (4 cm, 8 cm, 12 cm) with and without performing a secondary cognitive task at the same time (single-task vs. dual-task condition). The dependent variable is the distance subjects walk before stepping off the beam, averaged over three trials in each condition. The goal is to determine the diagnostic value of the beam tasks to predict falls in old age. The study protocol also measures a wide range of health-related indices and cognitive-processing related abilities. These measures are to be used as covariates to predict future falls of participants. The basic expectation is that the beam tasks carry added value for this prediction. In this supplement the focus is only on the dependent variable walking *Distance* in the experimental design comprising the between-subject factors *Age* group and *Sex* and the within-subject factors single vs. dual *Task* and *Width* of beam.

All analyses were done in the R environment [1]. For data processing and figures we used the *tidyverse* packages [2], the *afex* package [3], and the *ggplot* package [4]. R scripts used for the analyses in this supplement are considered as pre-registered analyses script for the final analyses and are available upon request.

Pilot study

Based on data in 20 Japanese young (10 female, M: 22 years , range: 19 to 25 years) and 16 old (9 female; M: 71 years, range: 66 to 77 years) healthy subjects [5], we derived hypotheses and an expected pattern of means. In this supplement we document the results of the pilot study and how we used them to arrive at estimates of statistical power for two theoretically relevant three-factor interactions, namely Age x Sex x Task and Age x Sex x Width.

Figure 1 displays the observed pattern of means of the Age(2) x Sex(2) x Task(2) x Width(3) design; errorbars are ±1 SE. There are clear ceiling effects for the beam width of 12 cm. Therefore, these data are no longer considered. A mixed-model ANOVA without the 12-cm beam width, yielded significant effects of Age, Task, and Width as well as significant Age x Task, Age x Width, and Sex x Width interactions – all of them in the expected canonical direction. The corresponding inferential statistics are shown in Table 1.

Table 1. Sources of variance and statistics for Age (2) X Sex (2) x Task (2) x Width (2) mixed-model ANOVA of pilot study.

Source of Variance	num Df	Error SS	den Df	F value	Pr(>F)	
Age	1	26.2	32	42.51	2.4e-07	***
Sex	1	26.2	32	0.80	0.3769	
Age:Sex	1	26.2	32	0.32	0.5760	
Task	1	12.6	32	16.47	0.0003	***
Age:Task	1	12.6	32	27.24	1.1e-05	***
Sex:Task	1	12.6	32	2.35	0.1349	
Age:Sex:Task	1	12.6	32	1.10	0.3018	
Width	1	20.2	32	142.93	2.4e-13	***
Age:Width	1	20.2	32	11.38	0.0020	**
Sex:Width	1	20.2	32	9.50	0.0042	**
Age:Sex:Width	1	20.2	32	0.44	0.5137	
Task:Width	1	12.6	32	0.34	0.5655	
Age:Task:Width	1	12.6	32	0.53	0.4710	
Sex:Task:Width	1	12.6	32	0.55	0.4629	
Age:Sex:Task:Width	1	12.6	32	0.04	0.8354	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

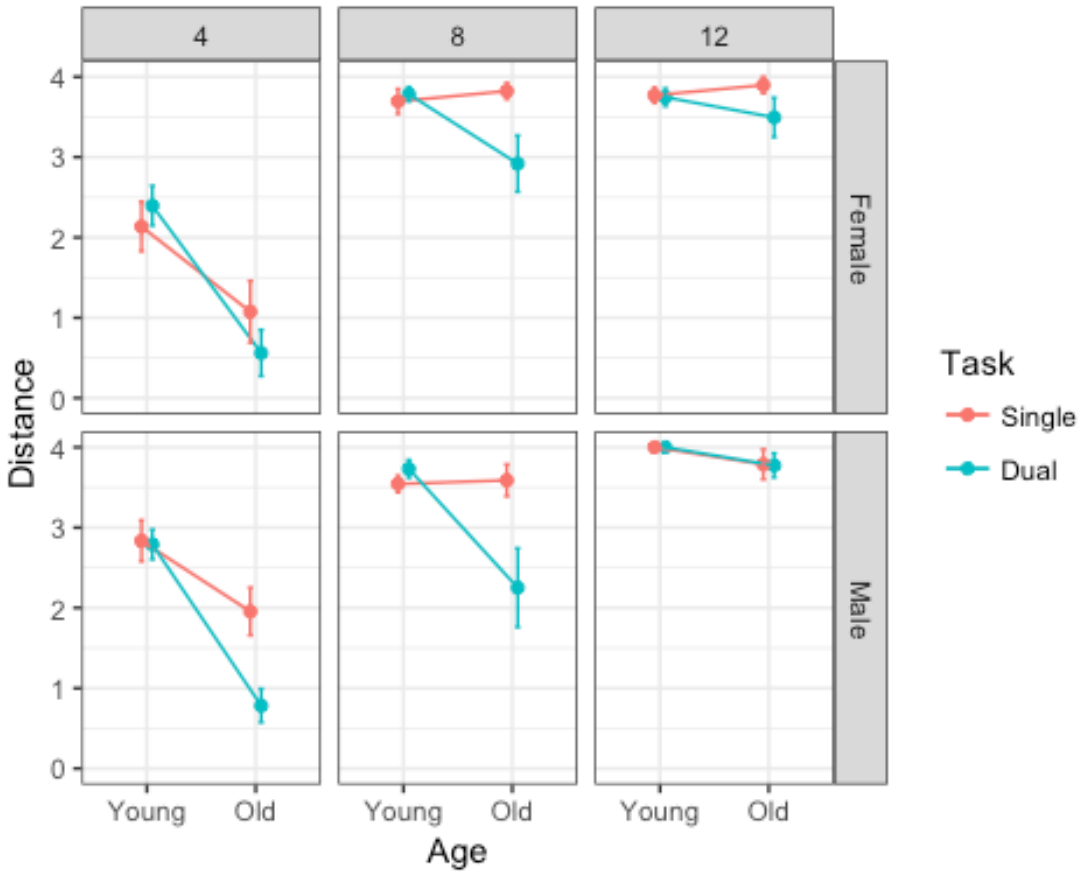

Figure 1. Means of Age (2) X Sex (2) x Task (2) x Width (2) design of pilot study. Errorbars are ±1 standard error of cell mean.

None of the higher-order interactions were significant, but two of them (i.e., Age x Sex x Task and Age x Sex x Width) exhibited theoretically relevant complementary sex-related patterns. They will serve as targets for the planned project. Specifically, old male subjects appear to be especially affected by the dual-task condition; old female subjects appear to be especially affected by the narrow 4-cm width of the beam (Figures 3, 4).

Extrapolation to full design

We expand the experimental design to seven age groups instead of two, covering the seven decades from 20 to 80+ years, that is Age (7) x Sex (2) x Task (2) x Width (2). The data of this experiment are completely determined by (a) the means and standard deviations of the 14 Age (7) x Sex (2) between-subject cells, (b) the correlations between 4 measures relating to the Task (2) x Width (2) within-subject conditions, and the assumption that residuals will be normally distributed. The R function mixedDesign(..., empirical=TRUE, ...), a wrapper for the mvrnorm function from the MASS package [6] allows one to simulate data from a multivariate normal distribution that exactly match the *a priori* specification [7].

Starting with rounded means of the pilot study (see Figure 1 and means for A1 and A7 in Figure 2), we extrapolated means and standard deviations for the intermediate five age groups A1 to A6 as shown in Figure 2, assuming that differences would increase with age. This figure represents our expectation (hypothesis) for the pattern of means corresponding to the four-factor interaction of the design. We will use this profile of means to determine the sample size needed for adequate statistical power (~80%) to detect significant 3-factor interactions of Age x Sex x Task and Age x Sex x Condition under different assumptions for standard deviations and correlations.

Expected means of full factorial design

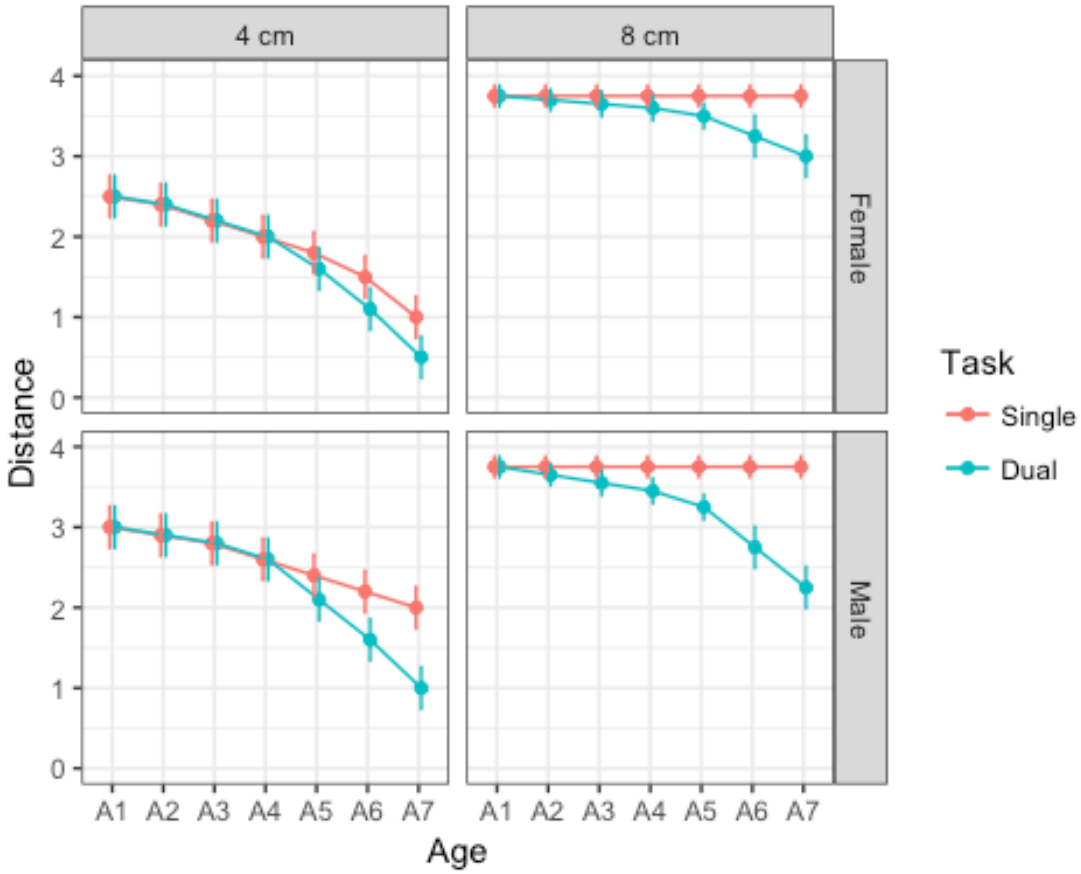

Figure 2. Extrapolated expected means of Age (7) X Sex (2) x Task (2) x Width (2) design for determination of statistical power. Errorbars are ±1 standard error of cell mean.

Three-factor interactions

Given the profile of means shown in Figure 2, we can also compute and visualize the two theoretically critical 3-factor interactions implied by this specification: (1) Age x Sex x Task and (2) Age x Sex x Width.

Age x Sex x Task

As shown in Figure 3, according to our *a priori* specification, age differences in the task effect increase more strongly for male than female participants.

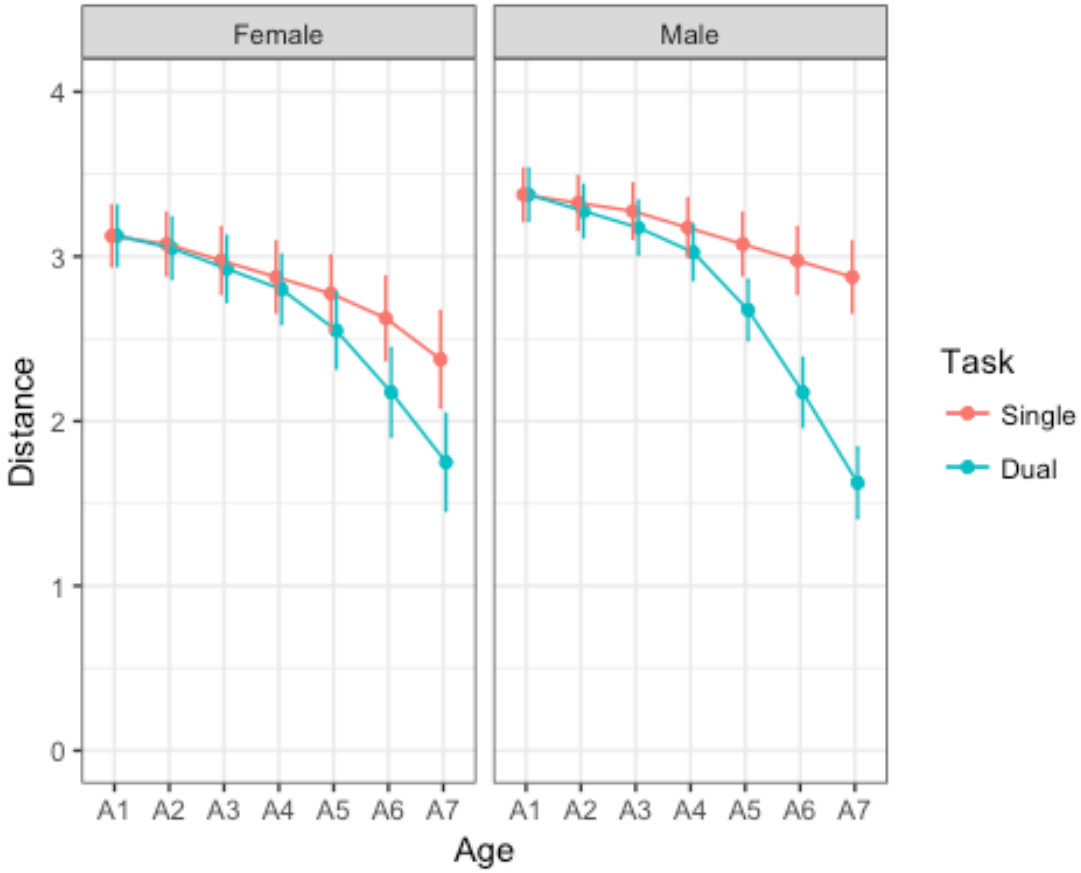

Figure 3. Extrapolated expected Age x Sex x Task interaction. Errorbars are +/- 1 standard error of cell mean. This interaction is one of two targets of the power simulation.

Age x Sex x Width

As shown in Figure 4, according to our *a priori* specification, age differences in the width effect increase more strongly for female than male participants.

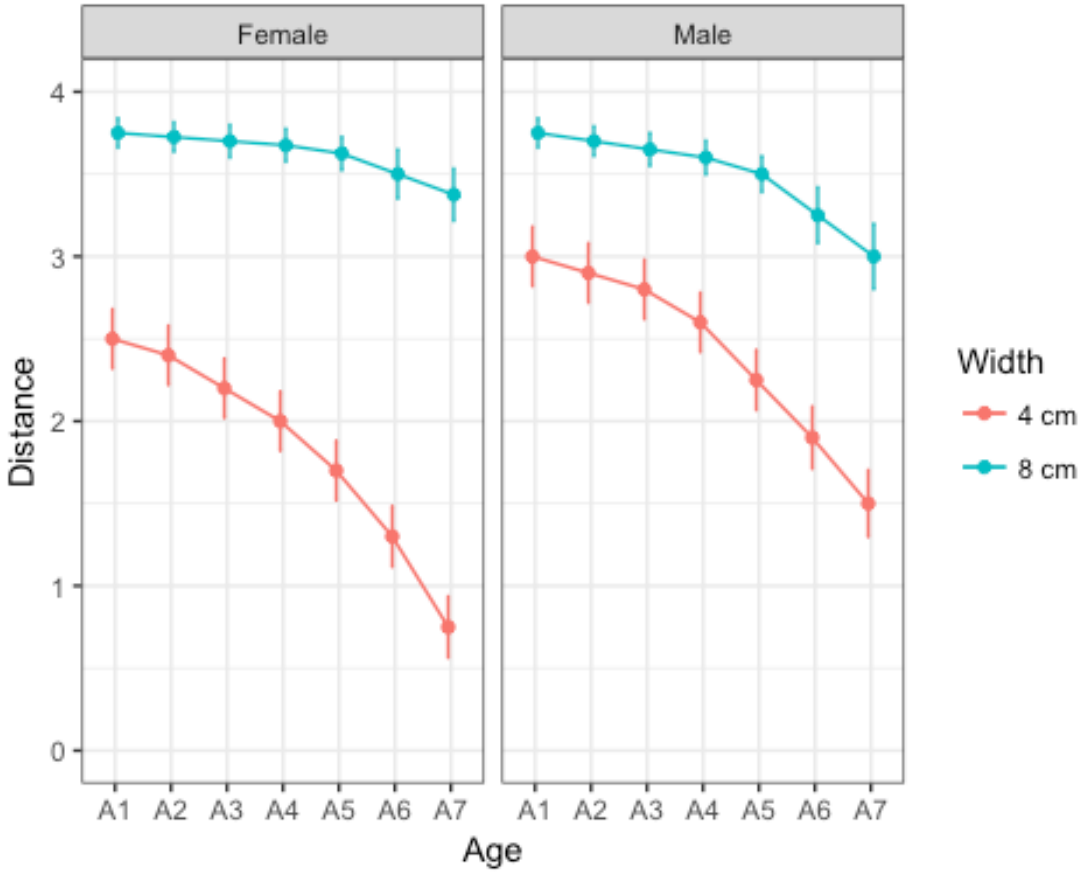

Figure 4. Extrapolated expected Age x Sex x Width interaction. Errorbars are +/- 1 standard error of cell mean. This interaction is one of two targets of the power simulation.

Power simulations

To determine statistical power for various scenarios, we simulate the experiment manifold using again the mixedDesign function. This time, however, we did not generate data that matched the specifications, but data that were drawn at random from a population with the specifications, that is mixedDesign(..., empirical=FALSE, ...). Specifically, we draw 500 random samples of an *a priori* specified size from a normal distribution with *a priori* specified population parameters representing expecected cell means and standard deviations as well as correlations within the cell means.

Each of the 500 samples was analyzed with linear mixed model (LMM) using the lmer()function of the *lme4* package [8]. We counted the number of significant effects for each source of

variance and divided this number by 500. This provided us with the power to detect the effect specified in the means of Figure 2. The results of these power analyses are summarized in Table 2 and will be described below.

Table 2. Results of power simulations

	Power (@ alpha = .05)			
	n = 15 r = .70 SD = 1	n = 20 r = .70 SD = 1	n = 20 r = .50 SD = 1	n = 20 r = .50 SD special
Source of variance				
Age.L	1.00	1.00	1.00	1.00
Age.Q	0.39	0.48	0.57	0.74
Sex	0.46	0.58	0.70	0.86
Task	1.00	1.00	1.00	1.00
Width	1.00	1.00	1.00	1.00
Age.L x Sex	0.05	0.07	0.07	0.07
Age.L x Task	1.00	1.00	1.00	1.00
Sex1 x Task	1.00	1.00	1.00	1.00
Age.L x Width	1.00	1.00	1.00	1.00
Sex x Width	0.70	0.78	0.62	0.72
Task x Width	0.62	0.73	0.54	0.66
Age.L x Sex x Task	0.67	0.78	0.56	0.68
Age.L x Sex x Width	0.68	0.82	0.60	0.74
Age.L x Task x Width	0.24	0.34	0.26	0.31
Sex x Task x Width	0.15	0.18	0.11	0.14
Age.L x Sex x Task x Width	0.10	0.11	0.06	0.08

Notes. Estimates are based on 500 simulations using the same matrix of means (see Figure 2). Dependent variable is distance walked on 4-m beam under 2 Width (4 cm, 8 cm) x 2 Task (single, dual) within-subject conditions. n is number of subjects in each of Age (7) x Sex (2) cells of between-subject design; Age.L = linear trend of Age, Age.Q = quadratic trend of Age. SD special: special matrix of standard deviations (if mean of Distance < 3.75, SD = 1, else 0.7. r is correlation between four measures from Task (2) x Width (2) conditions.

Simulations with statistical assumptions met

In this section, we determine the sample size needed for adequate statistical power for the two three-factor interactions assuming the profile of means of Figure 2 and that critical statistical assumptions are met. Note that the actual sample size will be largerThe assumptions comprise:

(1) balanced design (equal n in all cells)
(2) homogeneity of variance equal standard deviations in all cells

(3) sphericity (equal correlations in all cells)

In addition, we assumed that age trends are at most quadratic, that is we specified a linear and quadratic effect of age for the factor Age (see Figures 2 to 4). Preliminary analyses also revealed that statistical power is only realistic for interactions between the linear trend of age with the other factors. Therefore, as far as Age is concerned, we included main effects for linear and quadratic effects of age, but interactions only between the linear effect of Age and the other three factors.

A sample size of 15 men and 15 women (30 in each age group) represents a very realistic scenario of a balanced design for this study, given that we must expect an increasing dropout rate especially in the oldest age bracket. We assume a standard deviation of 1.0 and a correlation of .70.

The first column of Table 2 shows the probabilities that main effects and interactions are significant at the 5% level of error. Statistical power for the two critical interactions was .67 (Age x Sex x Task) and .68 (Age x Sex x Task), respectively.

The second column of Table 2 displays statistical power with 20 subject in each design cell (40 subjects in each age group). Statistical power for the two interactions increased to .78 and .81, respectively.

The third column of Table 2 shows that statistical power for the interactions was reduced if the correlation is not .70, but .50. They are now at .56 and .60, respectively.

Simulation with violation of statistical assumption

Statistical assumptions underlying analytical computation of statistical power are rarely realistic. Restriction of range in scores due to ceiling effects (e.g., 12-cm beam widths in single task conditions), age-group related differences in standard deviations (i.e., violation of assumption of homogeneity of variance) or in correlations between dependent measures (i.e., violation of assumption of sphericity) are difficult to take into account. Finally, it may be difficult to recruit enough participants for some of the cells of the design.

The power simulation used in this supplement allowed us to explore the consequences of varying assumptions about standard deviations in the between-subject cells of the design and about correlations between the within-subject measures. In principle, we could also explore the consequences of unbalanced designs.

As an example, we “violated” the assumption of equal standard deviations (SDs) as observed in the pilot data: For cells with a large mean (i.e. > 3.50), there was a restriction of range because the maximum value was 4 m. For all cells with the single-task condition this restriction resulted in a reduction of SDs from 1.0 to .50; for cells with the dual-task condition SDs depended on age (.50 for the two youngest age groups, .6 for the next three age groups, and 1.0 for the two oldest age groups). The differences in SDs are also shown in different errorbars in Figure 2). Running the simulations with these SDs yielded the statistical power estimates for the two 3-factor interactions displayed in the last column of Table 2 (i.e., .68 and .74, respectively).

When assumptions are violated one must check whether the violation changes the nominal alpha under the null hypothesis. This was not the case in this instance. Running the simulation under

the assumption of the null hypothesis (i.e., no differences between means) yielded 4.6% and 5% significant interactions.

References

1 R Development Core Team: R: A language and environment for statistical computing Vienna, R Foundation for Statistical Computing., 2018,

2 Wickham H, Grolemund G: R for data science. New York, O'Reilly, 2017.

3 Singmann H, Bolker B, Westfall J, Aust F: Analysis of factorial experiments. Vienna, R Foundation for Statistical Computing, 2018,

4 Wickham H: ggplot2: Elegant graphics for data analysis. . New York, Springer, 2009.

5 Uematsu A, Tsuchiya K, Yokei H, Suzuki S, Hortobágyi T: Cognitive dual-tasking augments age-differences in dynamic balance during beam walking. Exp Gerontol Submitted

6 Venables WN, Ripley BD: Modern applied statistics with S, ed 4th. New York, Springer, 2002.

7 Hohenstein S, Kliegl R: Simulation of factorial mixed-model designs in R: The mixedDesign() Function, 2012,

8 Bates D, Maechler M, Bolker B, Walker S: lme4: Linear mixed-effects models using Eigen and S4 Vienna, R Foundation for Statistical Computing, 2016,