Journal of Aging and Physical Activity, 2002, **10**, 143-159

Differences in Physical-Fitness Test Scores Between Actively and Passively Recruited Older Adults: Consequences for Norm-Based Classification

Marieke J.G. van Heuvelen, Martin Stevens, and Gertrudis I.J.M. Kempen

This study investigated differences in physical-fitness test scores between actively and passively recruited older adults and the consequences thereof for norm-based classification of individuals. Walking endurance, grip strength, hip flexibility, balance, manual dexterity, and reaction time were measured in participants age 57 years or older: 1 sample recruited through media announcements (passively recruited) and 1 sample recruited through personal contact (actively recruited). Classifications on a 5-point scale based on norms were cross-tabulated. Compared with the actively recruited sample, performance of the passively recruited sample was significantly better on all tests except, for women, hip flexibility and manual dexterity. Cross-tabulation of the 2 classifications showed that percentages of agreement varied from 27.4% to 87.4%. Cohen's Kappa varied from .11 to .84. Caution should be used when giving feedback on test performance and subsequently making physical activity recommendations if norms are based on the performance of passively recruited older adults.

Key Words: physical performance, physical-fitness standards, self-selection, volunteer bias, elderly

In recent years, interest in performance-based physical-fitness tests has increased (Lemmink, 1996; Osness et al., 1996; Rikli & Jones, 1999a). Physical fitness can be considered a multidimensional construct including components such as endurance, strength, flexibility, coordination, balance, and reaction time (Fleishman, 1964; Greene, Williams, Macera, & Carter, 1993; Marsh, 1993). Results from physical-fitness tests have been used to investigate average performance in groups of participants, for instance in epidemiological research on the aging process (e.g., van Heuvelen, Kempen, Ormel, & Rispens, 1998; Kozma, Stones, & Hannah, 1991; Rikli & Busch, 1986) and in physical activity intervention studies (e.g., Brown & Holloszy, 1991; Morey et al., 1989; Stevens, Bult,

van Heuvelen and Stevens are with the Institute for Human Movement Sciences, University of Groningen, 9700 AD Groningen, the Netherlands. Stevens is also with the Dept. of Orthopaedics, University Hospital of Groningen. Kempen is with the Dept. of Medical Sociology, University of Maastricht, 6200 HD Maastricht, the Netherlands.

de Greef, Lemmink, & Rispens, 1999). On the individual level, physical-fitness tests have been used to inform people about their physical-fitness status and to advise them about exercise and sports (Lemmink).

Physical-fitness tests can be either criterion- or norm-referenced (Baumgartner & Jackson, 1991). On criterion-referenced tests, participants are classified as either proficient or nonproficient, pass or fail. On norm-referenced tests, individual scores are compared with scores of norm populations, usually people of the same gender and age. For instance, the Groningen Fitness Test for the Elderly uses quintiles of a norm population to classify people on a five-point scale: *much worse, worse, comparable, better,* or *much better* than those of the same age and gender (Lemmink, 1996). Osness et al. classify participants on a three-point scale: *below average* (<1 *SD* of the mean), *average* (within 1 *SD* of the mean), and *above average* (>1 *SD* of the mean). Rikli and Jones (1999b) classified a large group of men and women who performed the Functional Fitness Test battery in 10th, 25th, 50th, 75th, and 90th percentiles. An individual's physical-fitness profile can be established on the basis of norm scores on a variety of physical-fitness tests. The physical-fitness profile provides information about less developed and more developed physical-fitness components and is used to give a personal physical activity recommendation.

Norms are developed using the test results of volunteers. These volunteers can be self-selected in such a way that they tend to be more fit than the general population. The amount of self-selection can depend on the method of recruitment used. Sarkin et al. (1998) distinguished passive and active recruitment methods. Passive recruitment begins with an invitation to participate, such as announcements in the media or mailed letters, but offers no further prompting by personal contact. Because it is not known who actually received the information and how the responders might differ from nonresponders, the participant pool is not clearly defined. The active-recruitment method begins with a defined participant pool from which participants are personally asked to participate, either by telephone or face to face. Actively recruited samples are expected to be less self-selected and thus more representative of the general population than are passively recruited samples (Sarkin et al.). For instance, Sarkin et al. showed that people who were recruited by direct mail for a health-promotion program reported more physical activity and lower blood pressures than those who were recruited by means of personal telephone calls. Nonetheless, the passive-recruitment method is more commonly used (Sarkin et al.).

To obtain norm scores of physical-fitness tests for older adults, Lemmink (1996) and Rikli and Jones (1999b) used passive methods to recruit a norm population. For instance, Rikli and Jones (1999b) used flyers posted in community centers and similar locations, media announcements (radio, newspapers, and television), and announcements made in various classes and meetings. These norm populations might therefore be healthier and fitter than the general population, and the norms might be too high. Consequently, feedback given to individual participants about their results might undervalue their real level of performance. Neither the amount of bias nor the variability in bias across different physical-fitness components has ever been investigated. Therefore, this article investigates differences in performance on a wide range of physical-fitness tests between passively and actively recruited samples of older adults and the consequences of these differences for norm-based classification of individuals.

Method

PARTICIPANTS

The method used in this study is a secondary analysis of data of two samples: a passive-recruitment sample (Lemmink, 1996) and an active-recruitment sample (van Heuvelen, 1999).

Passive-Recruitment Sample. By means of passive methods, a sample of older adults was recruited in five municipalities in the northern part of the Netherlands. Announcements were made by local welfare organizations, in local classes of the More Exercise for the Elderly National Foundation (Landelijke Stichting Meer Bewegen voor Ouderen), in local newspapers, on local radio stations, and on cable-TV information service. In the Netherlands, local media are well monitored by older adults. The announcements were directed at adults 55 years or older. Potential participants were given the chance to carry out performance-based physical-fitness tests, as well as have some health measures taken (blood pressure, respiratory functioning, and serum cholesterol). The suitability of the tests for all older adults, not only the fit and healthy ones, was emphasized. Furthermore, participants would get feedback on their test results and a personal recommendation for a healthier lifestyle. A total of 1,162 older adults participated in the physical-fitness tests. In the analyses, only those 57 years or older (N = 1,108) were included. The reason for this restriction was to enable a comparison with the active-recruitment sample, which also included people 57 years or older. Characteristics of the final sample are presented in Table 1.

Active-Recruitment Sample. The active-recruitment sample was selected in two steps: (a) drawing the sample of baseline participants for the Groningen Longitudinal Aging Study (GLAS) and (b) drawing the sample of participants for a physical-fitness study from the baseline participants. GLAS is a population-based prospective follow-up study of the determinants of health-related quality of

Table 1 Characteristics of the Samples, *M* (*SD*)

	Passive-Recruitment Sample		Active-Recruitment Sample	
	Men	Women	Men	Women
n	353	755	274	350
Age (years)	66.6 (5.8)	66.6 (6.3)	68.8 (7.6)	69.0 (7.7)
Body-mass index (kg/cm^2)	25.8 (2.4)	27.0 (3.5)	26.2 (2.7)	27.4 (3.9)
Diastolic blood pressure (mmHg)	86.6 (10.3)	86.1 (11.3)	86.6 (11.2)	85.7 (11.7)
Systolic blood pressure (mmHg)	147.3 (18.3)	149.0 (20.2)	150.6 (21.6)	152.7 (24.9)

life in older adults—in particular, physical and social disability and well-being (Kempen, Ormel, Brilman, & Relyveld, 1997; Ormel et al., 1997).

For the baseline sample, 27 general practitioners in the northern part of the Netherlands approached all their patients age 57 years or older (N = 8,723). The general practitioners requested permission to pass on names and addresses to the researchers. Of the total number approached, 1,937 people refused. When contacted by the researchers, another 1,277 people refused to participate and it was discovered that 152 had died or left the practice. Another 78 were excluded because of severe cognitive impairments (Mini Mental State Examination score <17; Folstein, Folstein, & McHugh, 1975). Ultimately, 62% participated in the baseline measurements, which included both face-to-face interviews during home visits and written questionnaires. In this first step, nonresponse was associated with age (34% for 57- to 69-year-olds, 42% for 70- to 84-year-olds, and 67% for adults 85 years or older) and to some extent with gender (37% for men and 41% for women). Furthermore, the proportion of patients with malignant neoplasms was higher among the nonresponders, but the same did not apply for heart disease, chronic respiratory disease, or chronic diseases of the musculoskeletal system (Kempen, Miedema, van den Bos, & Ormel, 1998).

In the second step, a random sample of the baseline participants (N = 770) was invited to participate in the physical-fitness study during the baseline face-to-face interview, of which 559 actually did (van Heuvelen, Kempen, Ormel, & de Greef, 1997; van Heuvelen et al., 1998). To overcome potential transportation barriers, participants were offered transport by car between the test location and their homes. Nevertheless, nonresponse in this second step was associated with perceived physical limitations: Nonparticipants had, on average, lower scores on the Physical Functioning Scale of the MOS Short-Form General Health Survey (Stewart, Hays, & Ware, 1988) than did participants. To compensate for this effect, an additional sample was included of 65 seniors with low scores (<34, which is equivalent to more than three physical limitations out of six) on the scale. Participants in the fitness study were representative of the baseline participants with respect to age and gender and only a little positively biased with respect to physical limitations and disability. Characteristics of this sample are presented in Table 1.

MEASURES

Physical fitness was assessed using performance-based tests of the Groningen Fitness Test for the Elderly (van Heuvelen et al., 1997, 1998; Lemmink, 1996). For both samples, the tests were administered in gymnasiums in participants' neighborhoods. The tests were as follows:

- *Walking endurance.* Participants walked on a rectangular course divided into three 16.7-m intervals. Walking speed was increased by 1 km/hr every 3 min, starting at a speed of 4 km/hr and ending at a speed of 7 km/hr. Participants had to keep up the effort as long as possible. The score was the number of completed intervals.
- *Flexibility of the hip and spine* (sit-and-reach test). Participants sat on the floor, legs outstretched, in front of a box. They then had to bend forward and push a slide over the box as far as possible with their fingertips. The shift of the slide was recorded. The best of three trials was taken as the obtained score.

- *Grip strength*. Participants held a handgrip dynamometer in the preferred hand with the arm at the side and squeezed it using maximum force. The score obtained was the best of three trials.
- *Balance*. Participants stood on a platform that could tilt sideways. For 30 s, they had to try to keep the platform in balance, without letting it touch the floor. The total time during this time interval that the platform was in balance was recorded. The final score was the best of three trials.
- *Manual dexterity*. Participants replaced 40 blocks from a full board to an empty board in a prescribed way as quickly as possible with the preferred hand. The time taken to complete the task was recorded.
- *Simple visual reaction time*. Participants reacted to a visual signal by pushing a button as quickly as possible. The time between signal and reaction was recorded. The score was the median of 15 trials.

Validity and reliability of the tests are described elsewhere (van Heuvelen et al., 1997, 1998). For all tests except manual dexterity and reaction time, higher scores indicated better performance.

DATA ANALYSIS

Because of positive skewness, manual dexterity and simple reaction time were transformed logarithmically. In addition, these tests were inverted so that higher scores indicated better performance. Subsequent analyses were stratified by sex and consisted of four steps.

In the first step, differences between the two samples in level of physical fitness were investigated with multivariate analysis of variance, with the six physical-fitness scores as dependent variables and sample (passive recruitment versus active recruitment) and age (three categories: 57–65 years, 66–74 years, and 75 years or older) as factors. The main effects for sample and age and the interaction effect Sample × Age were tested.

In the second step, for each fitness test, norms (quintiles) were calculated separately for both samples. The norms were controlled for age effects. First, linear-regression analyses were performed with the physical-fitness measures as dependent variables and age as the independent variable. Second, for each physical-fitness component, the 20th , 40th, 60th, and 80th percentiles of the residuals were computed. Next, for each age level, norm scores (quintiles of fitness scores) were calculated as the expected value for that age plus residuals' 20th, 40th, 60th, and 80th percentile.[1]

In the third step, each participant was classified on a five-point scale: *much worse than average performance* (score < 20th percentile), *worse than average performance* (20th percentile ≤ score < 40th percentile), *about average performance*

[1] An alternative way to compute norms is to calculate quintiles stratified by age category. Our approach has two related advantages. First, because when calculating norms for a certain age level, information of other ages has been incorporated, lower sample sizes are required to obtain accurate norm scores. Second, it is possible to calculate norms per age instead of per age category. For practical reasons, we did not present the raw norm scores. These norm scores are described in Lemmink (1996) for the passive-recruitment sample.

(40th percentile ≤ score < 60th percentile), *better than average performance* (60th percentile ≤ score < 80th percentile), and *much better than average performance* (score ≥ 80th percentile). The participants were classified for all fitness tests and according to the norms of both samples. For instance, a 74-year-old male participant from the passive-recruitment sample has a grip strength of 40 kilogramforce. According to his own group's norms his performance is worse than the average performance. According to the norms of the active-recruitment sample, however, this person's performance is classified as about average.

In the fourth step, the two classifications (based on norms of the active-recruitment sample and based on norms of the passive-recruitment sample) were cross-tabulated. Percentages of agreement, percentages of disagreement, and Cohen's Kappa were computed, and p values lower than .05 were considered to indicate statistical significance.

Results

Tables 2a and 2b show mean scores on the physical-fitness tests per age category and sex for the two samples separately. Both the multivariate results (Pillai's F = 22.4, df = 12,1212, $p < .001$ for men; Pillai's F = 35.4, df = 12,2136, $p < .001$ for women) and all univariate results (Table 3) showed significant main effects of age. Post hoc tests with Bonferroni corrections for multiple comparisons showed that the three age categories differed significantly for all tests except hip flexibility and reaction time between the youngest and the middle group for men. For all tests the youngest group performed best and the oldest group performed worst.

On average, the passive-recruitment sample performed better than the active-recruitment sample did on all tests. Results of the multivariate analysis of variance showed that the multivariate main effects of sample were significant for both men and women (Pillai's F = 15.3, df = 6,605, $p < .001$ for men; Pillai's F = 21.8, df = 6,1067, $p < .001$ for women). The univariate main effects of sample (Table 3) were significant for all tests except manual dexterity for both men and women and flexibility of the hip and spine for women.

Almost all test scores of the active-recruitment sample were more strongly related to age than were the test scores of the passive-recruitment sample. According to the multivariate analysis, this sample-by-age interaction effect was significant for both men and women (Pillai's F = 2.4, df = 12,1212, $p < .01$ for men; Pillai's F = 2.6, df = 12,1236, $p < .01$ for women). On the individual tests (Table 3), this effect remained significant for walking endurance, grip strength, balance, and reaction time for men and walking endurance, balance, and reaction time for women.

The consequences of the identified sample differences for norm-based participant classification are illustrated in the appendixes and in Table 4. Appendix 1 and Appendix 2 show the cross-tabulations in which the classification based on the norms of the passive-recruitment sample and the classification based on the norms of the active-recruitment sample were compared for men and women. Table 4 shows that percentages of participants who were classified in the same category according to both criteria vary from 27.4% for walking endurance in men to 87.4% for manual dexterity in women. Likewise, Kappa varies from .11 to .84. In cases of

Table 2a Test Scores for Men by Age Group, *M* (*SD*)

	57–65 years		66–74 years		≥75 years		Total	
	PR (*n* = 169)	AR (*n* = 108)	PR (*n* = 144)	AR (*n* = 110)	PR (*n* = 40)	AR (*n* = 56)	PR (*n* = 353)	AR (*n* = 274)
Walking endurance	60.1 (11.6)	53.1 (15.0)	54.5 (13.5)	40.8 (18.7)	41.5 (15.4)	26.5 (16.1)	55.7 (14.0)	42.7 (19.5)
Grip strength	48.4 (7.0)	47.4 (8.2)	44.3 (6.6)	42.8 (7.3)	40.0 (6.1)	34.3 (6.9)	45.8 (7.3)	42.8 (9.0)
Hip flexibility	25.5 (9.4)	22.9 (10.0)	23.0 (9.4)	22.3 (8.8)	21.1 (7.4)	19.0 (7.9)	24.0 (9.3)	21.9 (9.2)
Balance	80.1 (9.9)	78.8 (10.6)	76.6 (10.7)	73.7 (9.7)	73.5 (11.0)	63.9 (13.6)	77.9 (10.6)	73.8 (12.1)
Manual dexterity[a]	46.4 (5.0)	47.8 (9.7)	49.7 (6.2)	51.9 (12.9)	54.4 (5.8)	61.3 (17.4)	48.6 (6.1)	52.2 (13.7)
Reaction time[a]	221 (41)	231 (43)	225 (42)	241 (41)	237 (36)	280 (87)	224 (41)	245 (57)

Note. PR = passive-recruitment sample; AR = active-recruitment sample.
[a]Untransformed and uninverted scores.

Table 2b Test Scores for Women by Age Group, *M* (*SD*)

	57–65 years		66–74 years		≥75 years		Total	
	PR (*n* = 364)	AR (*n* = 131)	PR (*n* = 298)	AR (*n* = 137)	PR (*n* = 93)	AR (*n* = 82)	PR (*n* = 755)	AR (*n* = 350)
Walking endurance	49.4 (13.5)	43.8 (15.7)	40.9 (13.4)	31.4 (15.9)	30.4 (12.9)	16.2 (12.7)	43.7 (14.8)	32.5 (18.4)
Grip strength	31.3 (5.3)	28.9 (5.5)	28.5 (4.8)	26.8 (5.0)	25.4 (4.6)	22.2 (5.1)	29.4 (5.4)	26.5 (5.8)
Hip flexibility	31.6 (8.5)	31.9 (7.8)	30.2 (8.7)	29.6 (9.0)	28.2 (9.2)	25.6 (8.6)	30.6 (8.7)	29.5 (8.8)
Balance	78.9 (10.8)	77.7 (10.6)	74.7 (11.9)	72.1 (11.6)	70.7 (13.1)	60.1 (14.2)	76.2 (11.9)	71.5 (13.6)
Manual dexterity[a]	45.0 (5.2)	44.8 (4.8)	48.0 (6.4)	48.8 (9.5)	51.7 (7.8)	54.3 (10.2)	47.0 (6.5)	48.6 (9.0)
Reaction time[a]	232 (40)	248 (46)	246 (54)	257 (56)	249 (44)	292 (81.1)	240 (47)	262 (62)

Note. PR = passive-recruitment sample; AR = active-recruitment sample.

[a] Untransformed and uninverted scores.

Table 3 Results of Analysis of Variance for Men and Women

	Main effect sample (*F*, *df* = 2610[a]/1072[b])	Main effect age (*F*, *df* = 2610[a]/1072[b])	Interaction effect Sample × Age (*F*, *df* = 1610[a]/1072[b])
		Men	
Walking endurance	70.0***	81.9***	3.9*
Grip strength	15.2***	82.2***	4.6*
Hip flexibility	4.8*	7.0**	0.6
Balance	20.6***	36.4***	5.2**
Manual dexterity	6.8**	62.5***	1.4
Reaction time	27.5***	16.7***	3.2 *
		Women	
Walking endurance	75.4***	171.6***	5.6**
Grip strength	37.4***	95.8***	0.9
Hip flexibility	2.0	16.8***	1.4
Balance	32.1***	70.6***	9.7***
Manual dexterity	3.7	97.5***	2.5
Reaction time	39.0***	20.7***	3.5*

[a] *df* (error) for men; [b] *df* (error) for women.
*$p < .05$; **$p < .01$; ***$p < .001$.

disagreement, the norms of the passive-recruitment sample generally resulted in lower classifications than did the norms of the active-recruitment sample.

Discussion

This study showed a higher level of physical fitness in a sample of passively recruited older adults than in an actively recruited sample. These differences in physical fitness result in different norm-based classifications. Classification based on the norms of the passive-recruitment sample often leads to an undervaluation of the level of performance. This is an important finding that should be kept in mind when giving feedback about test performance and subsequently giving recommendations to individual participants. If norms are based on a sample of passively recruited volunteers, without a special effort having been made to get a representative sample, ratings such as "(much) worse than other people of your age and gender" should be given with the utmost caution.

The results of this study suggest that the oldest passively recruited volunteers self-selected more strongly than did the younger passively recruited volunteers. The Age × Sample interaction effect was significant, with stronger age differences in

Table 4 Comparison of Classification for Men and Women: Cohen's Kappa and Percentages of (Dis)agreement

	Cohen's Kappa	Percentage agreement	Percentage AR lower classification[a]	Percentage AR higher classification[b]
		Men		
Walking endurance	.111	27.4	0.0	72.6
Grip strength	.577	66.0	0.0	34.0
Hip flexibility	.658	72.6	0.0	27.4
Balance	.587	66.8	0.5	32.7
Manual dexterity	.779	82.3	6.7	11.0
Reaction time	.347	49.5	0.0	50.5
		Women		
Walking endurance	.196	34.8	0.0	34.8
Grip strength	.321	45.3	0.0	45.3
Hip flexibility	.808	84.7	3.4	11.9
Balance	.601	67.9	0.9	31.2
Manual dexterity	.843	87.4	7.2	5.4
Reaction time	.454	56.0	0.0	56.0

Note. AR = active-recruitment sample; PR = passive-recruitment sample.
[a]Norms of the AR sample lead to lower classification than norms of the PR sample do; [b]Norms of the AR sample lead to higher classification than norms of the PR sample do.

physical fitness for the active-recruitment sample. Consequently, when norms are based on a passive-recruitment sample, the older the participant the more likely he or she is to be classified in a lower class. Additional analyses showed that participants whose performance was undervalued using passive-recruitment norms (i.e., with a lower quintile ranking according to passive-recruitment norms than active-recruitment norms; see Table 4 and the appendixes) were, on average, older (69.1 years) than participants who were overvalued (66.6 years) or classified consistently (overall mean 66.5 years; averaged over the six physical-fitness tests, weighted with the number of participants who were undervalued). Where the oldest old are concerned, therefore, feedback on test performance should be given only when the potential bias resulting from self-selection has been taken into consideration.

These results raise the question of how the passively recruited volunteers selected themselves and why self-selection is strongest in the oldest participants. Age is positively related to health problems and disability (Jette, 1996). Consequently, with rising age, health problems and disability increasingly form barriers to getting to a gymnasium or participating in a physical-fitness test. People who were able to participate in a physical-fitness test but unable to come to a gymnasium were not represented in the passively recruited sample but did make up part of the actively recruited sample, which was offered transportation to the gymnasium.

The concept of experienced self-efficacy might provide another explanation. Self-efficacy is an important determinant for physical activity behavior (Dishman & Sallis, 1994), especially self-efficacy expectancies in one's abilities and for overcoming barriers to participation (Stevens, Bakker-van Dijk, de Greef, Lemmink, & Rispens, 2001). Self-efficacy expectancies for exercise decline with age (Clark, 1996). Therefore, of those still able to come to a gymnasium and participate in a physical-fitness test, those with lower levels of self-efficacy expectancies in the mentioned domains might prefer to decline participation when they are not given special encouragement.

Differences between classifications based on the passive- and active-recruitment samples vary across the physical-fitness components. For both men and women, the lowest percentages of agreement and the lowest Kappas were found for the walking-endurance test, the grip-strength test, and the reaction-time test. Only the manual-dexterity test (men and women), the flexibility test (men and women), and the balance test (women) passed the criterion of Kappa > .6 for acceptable Kappa (van de Sande, 1984). Consequently, the physical-fitness profile, which should provide information about the differences in performance on several physical-fitness components, will be biased if norms are based on volunteers who were recruited passively. In turn, this has consequences for exercise recommendations, which will tend to focus on improving endurance, strength, and reaction time. On the other hand, walking endurance and grip strength are the most important physical-fitness components for the performance of activities of daily living (ADLs; van Heuvelen, Kempen, Brouwer, & de Greef, 2000), so that in practice the bias does not work out negatively for the individuals concerned.

This raises the question of why the passively recruited volunteers' self-selection is so specific for certain physical-fitness components. Again, self-efficacy theory might provide part of the explanation. Self-efficacy expectancies depend on performance attainments (Bandura, 1982, 1997). In the physical domain, performance in ADLs and leisure-time physical activities will be most dominant. Because walking endurance and strength are most important for ADLs (van Heuvelen et al., 2000) and walking endurance is most closely related to level of physical activity (van Heuvelen et al., 1998), poorer functioning in these physical-fitness components might result in lower levels of physical self-efficacy expectancies in general and a stronger tendency to decline participation in a physical-fitness test.

This study calculated norms and classified participants with the purpose of illustrating the practical relevance of statistically significant differences in physical fitness between passively and actively recruited samples. Although we used our samples efficiently by calculating norms after controlling for age, much larger sample sizes are needed to obtain accurate norms for worldwide application.

This study might even underestimate self-selection of passively recruited individuals, especially for the oldest old. The active-recruitment sample was representative for the baseline sample of GLAS. Respondents of the baseline sample of GLAS, however, were self-selected with respect to age. Furthermore, institutionalized older adults were excluded a priori. Including institutionalized participants in the active-recruitment sample would result in larger differences between actively and passively recruited participants.

The results of this study demonstrate the necessity of exercising caution when using norm scores derived from test results of passively recruited samples. Ignoring self-selection could lead to misinterpretations and inaccurate physical activity recommendations.

Acknowledgments

We thank Dr. K.A.P.M. Lemmink for allowing us to use the results of the passive-recruitment sample. The results of the active-recruitment sample originate from the Groningen Longitudinal Aging Study (GLAS). GLAS is conducted by the Northern Centre for Healthcare Research (NCH) and various departments of the University of Groningen (RUG), the Netherlands. GLAS and its substudies are financially supported by the Dutch government (through NESTOR), the University of Groningen, the School of Medicine, the Dutch Cancer Foundation (NKB/KWF), and the Dutch Organization for Scientific Research (NWO).

References

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, **37**, 122-147.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Baumgartner, T.A., & Jackson, A.S. (1991). *Measurement for evaluation in physical education and exercise science.* Dubuque, IA: Brown.

Brown, M., & Holloszy, J.O. (1991). Effect of a low intensity exercise program on selected physical performance characteristics of 60-to-71-year olds. *Aging*, **3**, 129-139.

Clark, D.O. (1996). Age, socioeconomic status, and exercise self-efficacy. *Gerontologist*, **36**,157-164.

Dishman, R.K., & Sallis, J.F. (1994). Determinants and interventions for physical activity and exercise. In C. Bouchard, R.J. Shephard, & T. Stephens (Eds), *Physical activity, fitness, and health. International proceedings and consensus statement* (pp. 214-238). Champaign, IL: Human Kinetics.

Fleishman, E.A. (1964). *The structure and measurement of physical fitness.* Englewood Cliffs, NJ: Prentice-Hall.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). Mini Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189-198.

Greene, L.S., Williams, H.G., Macera, C.A., & Carter, J.S. (1993). Identifying dimensions of physical (motor) functional capacity in healthy older adults. *Journal of Aging and Health*, **5**, 163-178.

van Heuvelen, M.J.G. (1999). *Physical activity, physical fitness and disability in older persons.* Unpublished doctoral dissertation, University of Groningen, the Netherlands.

van Heuvelen, M.J.G., Kempen, G.I.J.M., Brouwer, W.H., & de Greef, M.H.G. (2000). Physical fitness related to disability in older persons. *Gerontology*, **46**, 333-341.

van Heuvelen, M.J.G., Kempen, G.I.J.M., Ormel, J., & de Greef, M.H.G. (1997). Self-reported physical fitness of older persons: A substitute for performance-based measures of physical fitness? *Journal of Aging and Physical Activity*, **5**, 298-310.

van Heuvelen, M.J.G., Kempen, G.I.J.M., Ormel, J., & Rispens, P. (1998). Physical fitness related to age and physical activity in older persons. *Medicine and Science in Sports and Exercise*, **30**, 434-441.

Jette, A.M. (1996). Disability trends and transitions. In R.H. Binstock & L.K. George (Eds), *Handbook of aging and the social sciences* (4th ed., pp. 94-116). San Diego, CA: Academic Press.

Kempen, G.I.J.M., Miedema, I., van den Bos, G.A.M., & Ormel, J. (1998). Relationship of domain-specific measures of health to perceived overall health among older subjects. *Journal of Clinical Epidemiology*, **51**, 11-18.

Kempen, G.I.J.M., Ormel, J., Brilman, E.I., & Relyveld, J. (1997). Adaptive responses among Dutch elderly: The impact of eight chronic conditions on health-related quality of life. *American Journal of Public Health*, **87**, 38-44.

Kozma, A., Stones, M.J., & Hannah, T.E. (1991). Age, activity, and physical performance: An evaluation of performance models. *Psychology and Aging*, **6**, 43-49.

Lemmink, K.A.P.M. (1996). *De Groninger fitheidstest voor ouderen, ontwikkeling van een meetinstrument (The Groningen fitness test for the elderly, development of a measurement instrument)*. Unpublished doctoral dissertation, University of Groningen, the Netherlands.

Marsh, H.W. (1993). The multidimensional structure of physical fitness: Invariance over gender and age. *Research Quarterly for Exercise and Sport*, **64**, 256-273.

Morey, M.C., Cowper, P.A., Feussner, J.R., DiPasquale, R.C., Crowley, G.M., Kitzman, D.W., & Sullivan, R.J., Jr. (1989). Evaluation of a supervised exercise program in a geriatric population. *Journal of the American Geriatrics Society*, **37**, 348-354.

Ormel, J., Kempen, G.I.J.M., Penninx, W.J.H., Brilman, E.I., Beekman, A.T.F., & van Sonderen, E. (1997). Chronic medical conditions and mental health in older people: Disability and psychosocial resources mediate specific mental health effects. *Psychological Medicine*, **27**, 1065-1077.

Osness, W.H., Adrian, M., Clark, B., Hoeger, W., Raab, D., & Wiswell, R. (1996). *Functional fitness assessment for adults over 60 years* (2nd ed.). Dubuque, IA: Kendall/Hunt.

Rikli, R., & Busch, S. (1986). Motor performance of women as a function of age and physical activity level. *Journal of Gerontology*, **41**, 645-649.

Rikli, R.R., & Jones, C.B. (1999a). Development and validation of a functional fitness test for community-residing older adults. *Journal of Aging and Physical Activity*, **7**, 129-161.

Rikli, R.R., & Jones, C.B. (1999b). Functional fitness normative scores for community-residing adults, ages 60–94. *Journal of Aging and Physical Activity*, **7**, 162-181.

van de Sande, J.P. (1984). *Gedragsobservatie. Een inleiding tot systematisch observeren.* Groningen, the Netherlands: Wolters-Noordhoff.

Sarkin, J.A., Marshall, S.J., Larson, K.A.,Calfas, K.J., & Sallis, J.F. (1998). A comparison of methods of recruitment to a health promotion program for university seniors. *Preventive Medicine*, **27**, 562-571.

Stevens, M., Bakker-van Dijk, A., de Greef, M.H.G., Lemmink, K.A.P.M., & Rispens P. (2001). A Dutch translation of a questionnaire assessing self-efficacy in leisure-time physical activity. *Journal of Aging and Physical Activity*, **9**(2), 223-232.

Stevens, M., Bult, P., de Greef, M.H.G., Lemmink, K.A.P.M., & Rispens, P. (1999). GALM: Stimulating physical activity in sedentary older adults. *Preventive Medicine*, **29**, 267-276.

Stewart, A.L., Hays, R.D., & Ware, J.E. (1988). The MOS short-form general health survey. Reliability and validity in a patient population. *Medical Care*, **26**, 724-735.

Appendix 1. Cross-Tabulation of Classifications for Men (in Bold Numbers of Agreement)

		Active-Recruitment Classification				
	Passive-recruitment classification	Much worse than mean	Worse than mean	About mean	Better than mean	Much better than mean
Walking endurance	much worse than mean	**74**	87	48	1	
	worse than mean			46	71	3
	about mean				82	20
	better than mean					96
	much better than mean					**97**
Grip strength	much worse than mean	**101**	46			
	worse than mean		**75**	68	2	
	about mean			**36**	77	
	better than mean				**89**	20
	much better than mean					**113**
Hip flexibility	much worse than mean	**108**	35			
	worse than mean		**82**	42		
	about mean			**63**	62	
	better than mean				**85**	31
	much better than mean					**112**

		Active-Recruitment Classification				
	Passive-recruitment classification	Much worse than mean	Worse than mean	About mean	Better than mean	Much better than mean
Balance	much worse than mean	**95**	42	1		
	worse than mean		**98**	38	2	
	about mean		3	**47**	75	4
	better than mean				**73**	41
	much better than mean					**102**
Manual dexterity	much worse than mean	**107**	28			
	worse than mean	1	**101**	16		
	about mean		15	**89**	18	
	better than mean			17	**104**	7
	much better than mean				9	**114**
Reaction time	much worse than mean	**89**	78	7		
	worse than mean		**16**	95	5	
	about mean			**48**	73	3
	better than mean				**53**	54
	much better than mean					**103**

Appendix 2. Cross-Tabulation of Classifications for Women (in Bold Numbers of Agreement)

		Active-Recruitment Classification				
	Passive-recruitment classification	Much worse than mean	Worse than mean	About mean	Better than mean	Much better than mean
Walking endurance	much worse than mean	**133**	145	17		
	worse than mean		**24**	141	73	
	about mean			**1**	185	11
	better than mean				**43**	144
	much better than mean					**181**
Grip strength	much worse than mean	**155**	105			
	worse than mean		**65**	173		
	about mean			**33**	179	
	better than mean				**57**	147
	much better than mean					**190**
Hip flexibility	much worse than mean	**206**	28			
	worse than mean	4	**178**	41		
	about mean		4	**176**	31	
	better than mean			13	**173**	30
	much better than mean				17	**195**

		Active-Recruitment Classification				
	Passive-recruitment classification	Much worse than mean	Worse than mean	About mean	Better than mean	Much better than mean
Balance	much worse than mean	**167**	81	1		
	worse than mean		**144**	62	3	
	about mean		10	**123**	84	10
	better than mean				**115**	96
	much better than mean					**186**
Manual dexterity	much worse than mean	**211**	9			
	worse than mean	15	**198**	7		
	about mean		33	**175**	14	
	better than mean			27	**175**	30
	much better than mean				4	**206**
Reaction time	much worse than mean	**150**	123	2		
	worse than mean		**61**	149	1	
	about mean			**95**	111	
	better than mean				**120**	100
	much better than mean					**193**