Examining Clinical Meaningfulness in Randomized Controlled Trials: Revisiting the Well Elderly II

David Schelly, Alisha Ohl

OBJECTIVE. Randomized controlled trial (RCT) interventions often rely on *p* values, where statistical significance is assumed to provide evidence of an intervention effect. This study provides a secondary data analysis of the Well Elderly II RCT using multiple approaches that examine clinical meaningfulness.

METHOD. We reanalyzed the Well Elderly II RCT using effect size, standard deviation, standard error of measurement, minimal difference, a fragility index, an assessment of poor scores at baseline, and an analysis with a small subgroup of participants removed.

RESULTS. Although some participants improved on several scales, most stayed the same, and a small subset declined. Omitting a small subgroup of participants led to nonsignificant *p* values.

CONCLUSION. There is evidence that disparities in baseline scores and regression to the mean may have created the appearance of an intervention effect. Our methods of considering clinical meaningfulness suggest improved approaches to analyzing RCT data.

Randomized controlled trials (RCTs) are consistently regarded as the gold standard of study design (e.g., Portney & Watkins, 2015). This reputation is not without merit: Compared with observational studies, in which selection effects are a concern, RCTs involve random assignment into groups so that group differences in unmeasured, lurking variables are eliminated. If groups are relatively equal at baseline, postintervention differences (including those from placebo effects) tend to be attributed to the intervention. Assuming random assignment was conducted effectively, the only remaining consideration is whether improvements are clinically meaningful.

The primary way of assessing the effectiveness of interventions is to rely on *p* values, in which *p* values are calculated for postintervention differences between groups, and statistical significance is assumed to provide evidence of an intervention effect. However, this approach has major problems. Vocal critics of this method (e.g., Cohen, 1992; Goodman, 1999) have expressed exasperation that *p* values continue to receive undue respect after decades of debate and virtually no remaining arguments in favor of their exclusive use. It seems that the respect for *p* values is a product of habit rather than appropriateness (Freeman, 1993).

The criticism relates to the essence of *p* values. In an experimental design comparing groups, *p* values quantify the probability of obtaining a group difference at least as extreme as that observed, given that the null hypothesis of no difference is true. A statistically significant difference between a treatment and a control group does not provide sufficient information about an intervention's effectiveness, clinical meaningfulness, or intervention effect size (Kraemer et al., 2003). It is noteworthy that small effects can become statistically significant as the sample size increases (see Carver, 1978; Lin et al., 2013); in addition, if statistical assumptions are not met, clinically meaningless results can achieve significance, and *p* values alone fail to consider this possibility.

With these concerns in mind, authors have long suggested alternatives to significance testing. Cohen (1962, 1990) spent his career arguing for the universal reporting of effect size, and he expressed dismay that his simple formula failed to catch on (Cohen, 1992). With regard to clinical work, a range of authors have discussed the concept of clinical meaningfulness. Jacobson et al. (1984) were not the first to argue that the focus should not be on group means but on people who may benefit from a treatment. The difficult task is to develop a method of defining and quantifying benefit (Jacobson & Truax, 1991).

Jaeschke et al. (1989) discussed the minimal clinically important difference, which attempts to quantify "the smallest difference in score in the domain of interest which patients perceive as beneficial and which would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management" (p. 408). This difference can be measured either cross-sectionally between patients, often referred to as the *minimally important difference*, or longitudinally within patients, referred to as the *minimally important change* (MIC; Beaton et al., 2001; de Vet et al., 2006; de Vet & Terwee, 2010). Ongoing discussions that distinguish between these concepts, although important, are beyond the scope of this article; we refer rather generically to the broader intervention aim of creating change that is clinically meaningful.

To measure or approximate clinical meaningfulness, one can use anchor-based or distribution-based methods (Guyatt et al., 2002; Wyrwich et al., 2005). Anchor-based methods rely on an independently assessed "anchor" that is used to determine cutoff points on a scale with otherwise unknown parameters. For example, in patients with cancer experiencing breakthrough pain, Farrar et al. (2000) first measured pain scores with several subjective numeric scales. Next, they defined clinically meaningful pain relief as the point at which patients decide not to take additional opioid medication. Finally, using medication use as an anchor, they determined clinically meaningful changes in the subjective scales.

Distribution-based methods, which rely on the sample distribution and inferential statistics, are commonly used when anchors are unavailable. One well-defended method is to use standard error of measurement (*SEM;* Rejas et al., 2008; Wyrwich et al., 1999). Unlike the standard error of the mean, *SEM* relies not on the sample size but on the reliability of the scale in question, thereby accounting for measurement error that could pass for clinically meaningful change. Others have argued that clinical meaningfulness can be defined simply as a change of 0.5 standard deviation (*SD*), which for a reliability of 0.75 is precisely 1 *SEM* (Norman et al., 2003). Finally, others have continued to promote the reporting of effect size, such as Cohen's *d* (Samsa et al., 1999). However, these distribution methods do not directly measure anything clinical (for a review, see Crosby et al., 2003), so many authors have used the term *minimum detectable change* (MDC) to indicate the minimum change required to exceed measurement error (de Vet & Terwee, 2010; Wyrwich et al., 1999).

Even in the presence of anchors, determining meaningful change is difficult, especially with subjective patient-report measures. For example, Terwee et al. (2010) showed that MIC values vary by method, and values determined by a given method vary across studies. Anchor-based measures, they pointed out, suffer from validity problems because they tend to be more strongly correlated with follow-up scores than baseline scores. Yet, anchor-based methods are perhaps universally preferred over distribution-based methods, and distribution-based methods are universally preferred over *p* values for determining clinical meaningfulness. Finally, there is broad agreement that distribution-based methods approximate clinical meaningfulness (e.g., Rejas et al., 2008), so they should at least be used until anchors are established (Turner et al., 2010).

Purpose and Significance

The purpose of this article is to provide an analysis of clinical meaningfulness by demonstrating the use of distribution-based methods on a well-known occupational therapy intervention: the Well Elderly II RCT, also known as Lifestyle Redesign (Jackson et al., 1998). The Well Elderly II RCT consisted of group and individual sessions to counsel older

adults about their daily activities and to develop preventive lifestyle changes that are thought to trickle down to improvements in physical health, mental health, and life satisfaction (see Jackson et al., 1998).

The Well Elderly intervention is well known outside of occupational therapy and has arguably been shown to be cost effective (Clark et al., 2012). It is regularly cited, with more than 250 citations on Google Scholar. Moreover, the intervention is being replicated elsewhere (e.g., Johansson & Björklund, 2016; Mountain et al., 2017) and has been included as a guideline for health promotion in older adults by the United Kingdom's National Institute for Health and Care Excellence. We randomly sampled 50 of the articles that cited Clark et al. (2012), and 32 (64%) cited the study as having been effective or showing that occupational therapy community interventions are effective.

The Well Elderly II RCT relied on a treatment-control crossover design, in which analysis of covariance (ANCOVA) was used to control for various baseline covariates between the treatment and control groups. Researchers reported statistically significant differences for pain, vitality, social functioning, mental health, and the mental health composite score on the Medical Outcomes Study Short-Form Health Survey (SF–36v2®; Ware, 2000); life satisfaction using the Life Satisfaction Index–Z (LSI–Z; Wood et al., 1969); and symptoms of depression using the Center for Epidemiologic Studies Depression (CES–D; Radloff, 1977) scale. The change scores and *p* values were mostly analogous using paired *t* tests, which we used to focus on improvements in the treatment group in the first stage of the study. We considered these same measures with the broader question of clinical meaningfulness, using several distribution-based approaches to reanalyze the data for each of the measures that obtained statistical significance in the original report.

Method

Participants

Four hundred sixty older adults participated in the Well Elderly II RCT. After randomization and attrition, data analysis for the treatment group included 187 participants who were primarily female (71.7%), ranged in age from 60 to 90 yr (mean [*M*] = 74.1, *SD* = 7.9), and were ethnically diverse. Approximately half of the participants had at least some college education (51.3%) and earned less than $11,999 annually (50.8%). A control group of 173 participants had similar characteristics. Participants were recruited from a graduated care retirement community (7.0%), 11 senior housing residences (46.0%), and 9 senior activity centers (47.0%).

Data and Analysis

Focusing on the seven scales that showed statistical significance in Clark et al. (2012), we used several approaches to examine clinical meaningfulness: effect size, 0.5 *SD*, 2 *SEM*s, and minimal difference (*MD*); a fragility index (*F* index); number of participants with poor scores at baseline; and an analysis with "extreme improvers" omitted. For 0.5 *SD*, 2 *SEM*s, and *MD*, we generated threshold values for each scale. Then, change scores were calculated as the difference between the baseline and posttest values for each participant in the treatment and control groups, and the threshold values were used to group participants as having declined, stayed the same, or improved in their score for each scale. For effect size, the fragility index, and the analysis with extreme improvers omitted, we conducted paired *t* tests using Stata (Version 14.2; Stata Corporation, College Station, TX). Each method is discussed in further detail.

Effect Size

Effect size quantifies the strength of the relationship between independent and dependent variables (Kraemer et al., 2003). Numerous measures of effect size exist (e.g., correlation coefficients, Cohen's *d,* and measures of risk potency). In this article, we discuss Cohen's *d* ($d = [m1 - m2]/\sigma$), which is computed by taking the difference between the means of two groups and dividing that difference by the pooled *SD*. Cohen's *d* ranges from minus to plus infinity, but *d* values much greater than 1 are uncommon. Cohen (1992) provided general guidelines for interpreting effect sizes, with cutoffs for small ($d = 0.2$), medium ($d = 0.5$), and large ($d = 0.8$) effects. A medium effect is intended to be a change visible to

the naked eye of the careful observer, whereas a small effect is noticeably smaller but not trivial (Cohen, 1992). Although effect size provides information about the relative strength of statistically significant findings, Kraemer et al. (2003) noted that effect size values are relative and not readily interpretable in terms of how much people are affected by treatment.

Standard Deviation

SD is a measure of the amount of variation, or spread, in a set of values. The variance is calculated by summing the squared differences between each value and the mean, divided by the sample size. *SD* is the square root of the variance. To determine clinical meaningfulness, a value of 0.5 *SD* has been found to correspond with patient-reported minimal change across a variety of studies (Norman et al., 2003).

Standard Error of Measurement

SEM is an estimate of the reliability of an obtained score and is typically determined during assessment development. *SEM* is calculated by subtracting the test reliability (*r* coefficient or intraclass correlation coefficient) from 1, taking the square root of that difference, and multiplying the square root by the *SD* of the test scores ($SEM = S_x\sqrt{1-r}$). Clinically, *SEM* is typically used to create a band or confidence interval (CI) around an obtained score to arrive at a sense of the true score ($X \pm SEM$; Harvill, 1991), in which 1 *SEM* corresponds to a 68% CI, and 2 *SEM*s correspond to a 95% CI. Some authors have suggested using 1 *SEM* as a cutoff for meaningful aggregate changes (e.g., Copay et al., 2007), but many authors (including the SF–36 authors; Ware et al., 1994) deem 1 *SEM* to be too liberal for individual changes. Thus, we used 2 *SEM*s for our analysis.

Minimal Difference

Weir (2005) proposed using *SEM* to determine the *MD* (others have referred to this as *MDC;* see Busija et al., 2008) that approximates a real treatment effect when performing pre- and posttests (see also Beaton et al., 2001). *MD* is calculated by multiplying the *SEM* of the measure by 1.96 (the *z* score associated with a 95% CI) and, to account for error coming from two scores (pre and post) rather than one, the square root of 2 ($MD = SEM \times 1.96 \times \sqrt{2}$).

Fragility Index

Walsh et al. (2014) proposed a fragility index to quantify how "fragile" an RCT's results are. They calculated the index primarily for dichotomous outcomes, in which *p*-value calculations involve comparing the proportion of patients experiencing an outcome or event in a treatment group with the proportion experiencing the outcome or event in a comparison group. To calculate the fragility index in this scenario, Walsh et al. iteratively removed participants in the treatment group who experienced the outcome or event, and each time they removed a participant, they recalculated the *p* value, continuing until it exceeded .05. The required number of participants to lose statistical significance was the fragility index, in which lower numbers (e.g., <5) indicate that the significant result depends on very few outcomes.

Our fragility index (Ohl & Schelly, 2017) was calculated by iteratively omitting the most extreme improving participant from the *p*-value calculation (i.e., using paired *t* tests) until the *p* value exceeded .05. A value of 1, then, indicates that without the most extreme improver, the result would not obtain statistical significance.

Poor Scores at Baseline

Participants with poor scores at baseline are those in the undesirable extremes, such as those reporting severe pain (low scores) or depression (high scores). These people are the most likely to show improvement over time because many have nowhere to go but to improve. One concern is that although aggregate scores may show no statistically significant differences at baseline—which was the case in the Well Elderly II data—disproportionate extreme values

can effectively create consequential differences between the treatment and control groups. The 5th percentile of the combined treatment and control groups was used as the cutoff for poor scores for each scale (the 95th percentile was used for depression, in which higher scores indicate depressive symptoms). For example, for the Pain scale, this included people who reported "very severe" pain and either "extreme" or "quite a bit" of interference with daily activities and people who reported "severe" pain and "extreme" interference.

Extreme Improvers Omitted

To determine whether improvers on individual scales were the same people across many scales, as opposed to many people improving on only one or two scales, we used Excel (Microsoft Corp., Redmond, WA) to mark 1-*SEM* improvers on all seven scales. Participants in the treatment group who improved by at least 1 *SEM* on six or seven of the scales (n = 8) were labeled "extreme improvers" and were omitted for recalculations of paired t tests.

Measures

Medical Outcomes Study Short-Form Health Survey.

The SF–36v2 (Ware, 2000) is a health survey that contains 36 items and yields eight profile scores indicating various aspects of functional health and well-being. Four of the profile scores (physical function, role physical, bodily pain, and general health) yield a physical health composite score, and the remaining four (mental health, role emotional, social functioning, and vitality) yield a mental health composite score. The test–retest reliability coefficients of the profile scores are all equal to or greater than .80, with the exception of social functioning (r = .76; Ware, 2000). Reliability estimates for the physical and mental health composite scores typically exceed .90. The *SEM*, with a 95% CI, has been reported extensively in the literature and ranges from ±6–7 points for the composite scores to ±13–32 points for the profile scores (Ware, 2000).

Life Satisfaction Index–Z.

The LSI–Z (Wood et al., 1969) is a 13-item self-report measure of subjective well-being in the older adult population. Each item consists of a 3-point scale coded as 0 (*disagree*), 1 (*unsure*), or 2 (*agree*). Total scores range from 0 to 26, with higher scores indicating higher life satisfaction. Test–retest reliability is acceptable (r = .79).

Center for Epidemiologic Studies Depression Scale.

The CES–D scale (Radloff, 1977) is a 20-item self-report scale designed to measure depressive symptomatology in the general population. Respondents rate the frequency with which depressive symptoms have occurred over the past week using a 4-point scale coded as 0 (*rarely or none of the time*), 1 (*some or a little of the time*), 2 (*occasionally or a moderate amount of time*), or 3 (*most or all of the time*). Total scores range from 0 to 60, with higher scores indicating more depressive symptoms. During its development, the CES–D scale demonstrated high internal consistency in both general (α = .85) and clinical (α = .90) populations (Radloff, 1977). Test–retest reliability was moderate (rs = .45–.70), which the test developer attributed to the variable nature of depressive symptoms and methodological flaws with the consistency and method of data collection between pre- and posttests.

Results

Table 1 reports baseline and posttest means for the seven scales, and p values are included first as they were reported in Clark et al.'s (2012) original report. Clark et al. used ANCOVA to compare regression slopes between the treatment and control groups, in which the posttest score was predicted with several factors, including the baseline score. This method is certainly justifiable, but there is reason to believe it may be misleading under certain conditions of baseline score imbalances (see Lord, 1967). Thus, the second p values reported in Table 1 show the results of

Table 1. Changes Between Baseline and Posttest for Treatment (n = 187) and Control Group (n = 173) Participants and p Values Using Three Methods in the Well Elderly II Trial

Outcome	Baseline (*SD*)	Posttest (*SD*)	*p*		
			Clark et al. (2012)[a]	Independent-Samples *t* Test[a]	Paired-Samples *t* Test[b]
SF–36v2					
Pain					
Treatment	42.75 (11.60)	44.62 (11.20)	.02*	.019*	.003**
Control	44.53 (11.30)	44.38 (11.86)			.834
Vitality					
Treatment	50.02 (9.90)	51.29 (9.85)	.03*	.041*	.047*
Control	49.99 (9.46)	49.60 (11.23)			.588
Social functioning					
Treatment	44.54 (11.75)	45.36 (11.37)	.04*	.011*	.219
Control	46.57 (9.75)	45.00 (11.33)			.050*
Mental health					
Treatment	47.75 (11.89)	49.07 (10.70)	.03*	.066	.053
Control	47.44 (11.24)	47.16 (11.81)			.738
Mental composite					
Treatment	47.41 (11.80)	48.64 (10.63)	.03*	.026*	.074
Control	48.24 (10.47)	47.45 (12.01)			.318
LSI–Z					
Treatment	17.23 (5.68)	18.00 (5.37)	.03*	.064	.012*
Control	16.80 (5.71)	16.85 (5.49)			.900
CES–D scale					
Treatment	13.82 (10.81)	12.47 (9.68)	.03*	.017*	.028*
Control	12.97 (10.54)	13.53 (11.17)			.398

Note. *p* values shown were published in Clark et al. (2012) using analysis of covariance, using independent samples *t* tests, and using paired *t* tests. CES–D scale = Center for Epidemiologic Studies Depression scale; LSI–Z = Life Satisfaction Index–Z; *SD* = standard deviation; SF–36v2 = Medical Outcomes Study Short-Form Health Survey.
[a]One-sided hypothesis tests were used. [b]Two-sided hypothesis tests were used for a more conservative test and to detect possible declines.
*p < .05. **p < .01.

independent-samples *t* tests, which simply compare the differences between baseline and posttest without controlling for any factors. We present these *p* values to show that a simpler comparison of means produces very similar results. Only mental health and the LSI–Z move slightly beyond the typical .05 cutoff for statistical significance.

Table 1 also reports *p* values from paired-samples *t* tests, which are later used to consider changes without the extreme improvers (effect sizes are also calculated from the within-group changes). Note that Clark et al. (2012) reported *p* values for one-sided tests, but we agree with suggestions to use two-sided tests (see Bland & Altman, 1996). Although the one- versus two-sided debate is far from settled, most researchers seem to agree that two-sided tests should be the norm (Ringwalt et al., 2011), not least because they set a higher bar for significance, thereby protecting slightly against Type 1 errors. In this case, they also allow for declines between baseline and posttest, which are hypothesized to occur in this population in the absence of intervention (Jackson et al., 1998). Using two-sided tests, social functioning does show a significant decline (p = .050) in the control group. The treatment group shows significant improvements in pain (p = .003), vitality (p = .047), the LSI–Z (p = .012), and the CES–D scale (p = .028) but not in social functioning, mental health, or the mental health composite score.

Effect Size

Effect sizes are shown in the Cohen's *d* column in Table 2. In the treatment group, effect sizes ranged from 0.13 to 0.19, all less than the cutoff for small effects suggested by Cohen (i.e., d = 0.20). Pain (d = 0.16) and the CES–D scale

Table 2. Several Distribution-Based Indices of Clinical Meaningfulness for Treatment (n = 187) and Control Group (n = 173) Participants in the Well Elderly II Trial

Scale	Cohen's *d*	0.5 *SD*[a]			2 *SEMs*[b]			*MD*,[c] % Better	*F* Index	Poor at Baseline	Minus Improvers (n = 8), p
		% Worse	% Same	% Better	% Worse	% Same	% Better				
SF–36v2											
Pain											
Treatment	0.16	4.8	83.4	11.8	1.1	90.4	8.6	2.7	7	15	.028
Control		9.2	80.3	10.4	5.8	87.9	6.4	2.3		5	
Vitality											
Treatment	0.13	7.5	80.2	12.3	3.2	89.8	7.0	4.3	1	19	.331
Control		12.2	77.3	10.5	7.6	86.6	5.8	1.2		11	
Social functioning											
Treatment		4.8	88.8	6.4	0.5	97.9	1.6	0.0	0	31	.638
Control	–0.15	11.6	82.7	5.8	4.6	95.4	0.0	0.0		13	
Mental health											
Treatment		8.0	78.1	13.9	4.8	85.0	10.2	5.9	0	11	.425
Control		15.1	73.8	11.0	9.9	82.6	7.6	2.9		9	
Mental composite											
Treatment		23.0	47.1	29.9	16.0	62.0	21.9	15.5	0	13	.662
Control		26.2	48.3	25.6	22.1	55.8	22.1	12.2		5	
LSI–Z											
Treatment	0.14	5.9	82.9	11.2	5.9	82.9	11.2	5.3	4	9	.068
Control		12.8	76.2	11.0	12.8	76.2	11.0	4.1		10	
CES–D scale											
Treatment	0.19	7.5	77.0	15.5	5.3	81.8	12.8	7.0	2	11	.240
Control		15.0	74.0	11.0	13.3	78.0	8.7	5.2		8	

Note. Values indicate percentages of participants who scored worse, the same, or better using 0.5 standard deviation (*SD*) and 2 standard errors of measurement (*SEM*s) and better using minimal difference (*MD*); a fragility index; number of participants with poor scores at baseline; and p values with 8 extreme improvers omitted. CES–D scale = Center for Epidemiologic Studies Depression scale; LSI–Z = Life Satisfaction Index–Z; SF–36v2 = Medical Outcomes Study Short-Form Health Survey.

[a]One-half *SD* was 11.85, 10.45, 11.35, 9.05, 5.00, 5.68, and 8.58, in order of the table. Note that the value for the LSI–Z was obtained using the baseline *SD* from the sample because it has not been reported in the literature. [b]The 2 *SEM*s value was 15.0, 15.6, 25.7, 14.0, 6.3, 5.2, and 9.4, in order of the table. [c]*MD* (calculated as $SEM * 1.96 * \sqrt{2}$) was 20.79, 21.62, 35.62, 19.40, 8.73, 7.21, and 13.08, in order of the table.

(d = 0.19) were the largest effects. In the control group, the effect size for social functioning (d = –0.15), in which there was significant decline, was also below the cutoff for a small effect.

0.5 Standard Deviation

The 0.5-*SD* method, also shown Table 2, was the least conservative method for grouping participants as having declined, stayed the same, or improved. With the exception of the mental composite score, which has a low *SD* and therefore displays high numbers of improvers and decliners in the treatment and control groups, the percentages of improvers and decliners were relatively similar between the treatment and control groups. The percentages of improvers ranged from 6.4% to 15.5% in the treatment group and from 5.8% to 11.0% in the control group; the percentages of decliners ranged from 4.8% to 8.0% in the treatment group and from 9.2% to 15.1% in the control group. For the mental composite score, slightly more control group participants remained the same relative to the treatment group (48.3% vs. 47.1%); however, for the remaining scales, the treatment group had more participants who remained stable compared with the control group, with between 77.0% and 88.8% showing no improvement and no decline. Overall, the control group tended to have slightly more decliners than improvers, whereas the treatment group had between 3 and 15 additional improvers relative to decliners.

Standard Error of the Mean

The 2-*SEM*s method was more conservative than the 0.5-*SD* method. Again, with the exception of the mental composite score, the percentages of improvers ranged from 1.6% to 12.8% in the treatment group and from 0% to 11.0% in the control group; the percentages of decliners ranged from 0.5% to 5.9% in the treatment group and from 4.6% to 13.3% in the control group. The treatment group had more participants who remained stable compared with the control group for all scales, with 62.0% showing no improvement and no decline in the mental composite, and between 81.8% and 97.9% showing no improvement and no decline in the remaining scales. Overall, the control group again tended to have slightly more decliners than improvers, and the treatment group had between 2 and 14 additional improvers relative to decliners.

Minimal Difference

The *MD* method yielded the most conservative cutoff for clinical meaningfulness. The mental composite continued to be an exception because 15.5% of treatment participants improved (12.2% of control participants), compared with between 0% and 7.0% on the other scales; the control group had between 0% and 5.2% showing improvement on the other scales.

Fragility Index

The fragility index (see Table 2) shows that very few extreme improvers need to be omitted—from 1 to 7—for each scale to lose statistical significance. Three scales had fragility scores of 0 because the *p* values were nonsignificant with two-sided paired *t* tests instead of Clark et al.'s (2012) use of one-sided ANCOVA tests.

Poor Scores at Baseline

With the exception of the LSI–Z, more participants in the treatment group than in the control group displayed poor scores at baseline (see Table 2). Pain, vitality, social functioning, and the mental composite had substantially more participants with poor scores at baseline, with 8–18 additional participants with poor scores in the treatment group compared with the control group.

Extreme Improvers Omitted

Eight participants stood out as showing consistent improvement across the scales, showing at least 1 *SEM* improvement on six or seven of the reported scales. It is noteworthy that these participants also scored more poorly at baseline: As a group, they displayed worse scores by between 0.94 and 5.65 *SEM*s compared with the remaining group (combined control and treatment); their scores were worse by more than 2 *SEM*s on mental health, the mental composite score, the LSI–Z, and the CES–D scale. Table 2 shows an analysis of the treatment group in its entirety compared with the treatment group with these 8 participants omitted. The result is that the aggregate improvements from baseline to posttest all but disappear in all of the scales except for pain and the LSI–Z, and only pain remains statistically significant with a two-sided test.

Discussion

We reexamined the Well Elderly II RCT using distribution-based methods to provide an example of how to analyze and interpret data through the lens of clinical meaningfulness. The effect sizes for all statistically significant results were minimal to small, suggesting minimal aggregate improvements in the treatment group's perceptions of quality of life, life satisfaction, and depression. Clinically, small effect sizes may or may not be detectable to an occupational therapy practitioner or the participants. To provide a more thorough understanding of intervention effects than *p* values alone,

we conducted subsequent analyses that took into account measurement error (i.e., *SD*, *SEM*, *MD*), a fragility index, an assessment of poor scores at baseline, and an analysis with 8 participants with extreme improvements omitted.

SD, *SEM*, and *MD* were used to determine the proportion of participants with scores that stayed the same, improved, and declined. A consistent pattern presented across the three methods. Compared with the control group, a greater proportion of the treatment group improved, a greater proportion stayed the same (with the exception of the mental health composite), and a slightly smaller proportion declined. More important, the clear majority of participants in the treatment and control groups experienced no meaningful change on each scale.

Using the more conservative *MD*, we found little difference between groups: For example, on the Pain scale, 2.7% of the treatment participants compared with 2.3% of the control group participants improved. This small difference could be explained by baseline disparities between groups. Another way of thinking about the *MD* results is that the vast majority of both groups either stayed the same or declined in pain (97.3% in the treatment group and 97.7% in the control group). Considering measurement error in this way potentially provides alternative explanations for statistically significant differences between groups. Moreover, it provides a way to single out a small subset of the total sample for additional analyses, with the potential to identify characteristics in the improvers that might explain their response to the intervention.

Given that more participants stayed the same in the treatment group compared with the control group, it is possible that the intervention prevented decline; however, several counterpoints should be noted. First, the assessments with statistically significant improvements in the Well Elderly II RCT were all self-report or subjective measures of health, and they do not indicate any objective cognitive or physical functioning. Therefore, relatively stable scores between pre- and posttests may not translate into a maintenance of function because they may only indicate a maintenance in perception of function, and perceptions may be susceptible to a placebo effect. Second, in a population in which Lifestyle Redesign would be used, it is not clear how much decline to expect in a 6-mo period. To assess age-related declines, researchers conducting medical and epidemiological studies typically collect prospective longitudinal data spanning years to decades, using a combination of objective and subjective assessments (e.g., Inzitari et al., 2007; Paterniti et al., 2002; van Gelder et al., 2004). Considering how declines are typically assessed, the study design used in the Well Elderly II RCT may not be appropriate for examining age-related declines.

The additional analyses point to a small group of participants who are driving the effects, and these effects are at least partially a product of regression to the mean. In brief, regression to the mean is a universal feature of repeated measures in the presence of measurement error. To illustrate, consider the SF–36v2 Pain scale, in which two questions are combined to form a raw score that ranges from 2 to 11. Now take any group with a particular score in the middle, such as all participants who report a 5 at baseline. At posttest, some will go up, and some will go down, such that measurement error will pull at the baseline score from the top and the bottom. However, participants at 11 and those at 2, when measured again at posttest, are affected by measurement error unidirectionally, in which any change at all will pull the group mean down or up, respectively.

With regression to the mean in mind, the small group of participants with extreme values are of special concern. Specifically, the fragility index ranged from 1 to 7, suggesting that the statistical significance was vulnerable to a few extreme improvers. Next, 8 participants showed improvement on all or all but one of the seven scales, and these participants had worse scores (e.g., more pain, less satisfaction) by 0.94–5.65 *SEM*s at baseline compared with the remaining group, possibly because they had experienced a recent injury or negative life event before the intervention. Finally, there were concerning disparities in baseline scores for participants in these poor extremes: The treatment group had substantially more participants in the poor extremes—8 or more for the majority of the scales—compared with the control group. The implication is that because participants at the poor extremes at baseline are more likely to improve than others (regression to the mean alone will lead to more improvement, on average), the improvements in the treatment group—even in the absence of improvement in the control group—may be related to these disparities. It

should be noted that the design of the scales may even exaggerate this effect: When baseline distributions are skewed such that the scale midpoints are above the means and few participants report the poor extremes, as is the case for the reported scales, regression to the mean will be asymmetrical, resulting in an aggregate improvement. Interestingly, the developer of the CES–D scale warned about this possibility (Radloff, 1977, p. 391).

In addition, the intervention may have had a placebo effect on participants who had poor scores at baseline. For example, if 2 participants had hip replacement surgery immediately before the intervention, they both would have been primed to improve on various measures, such as pain, after their surgery. If 1 participant was treated and 1 participant was placed in the control group, the treated participant may have reported slightly greater improvements than the participant in the control group simply because he or she was aware of being treated. Likewise, the individual sessions of the intervention may have targeted his or her rehabilitation. If cases such as this example explain the small group of improvers, then the effectiveness of Lifestyle Redesign may stem from rather traditional occupational therapy rather than preventive strategies.

Limitations

Although this article provides a more detailed analysis of the statistically significant findings in the participants and provides a discussion of the intervention's clinical meaningfulness, it does not include all possible methods of examining clinical meaningfulness, and it does not elucidate the underlying mechanisms driving the improvements in the participants.

Implications for Occupational Therapy Practice

Distribution-based approaches yield a more nuanced tableau than *p* values alone. These approaches suggest that the Well Elderly II intervention was, at best, beneficial for a select group. At worst, the small aggregate improvements could be explained by a combination of regression to the mean and a placebo effect. These findings may be indicative of an endemic pattern in RCTs in which the *p*-value standard provides somewhat of a blinder to true intervention effects. With this in mind, we offer the following recommendations for occupational therapists and researchers:

- For evidence-based practice, RCTs that only report *p* values should not be regarded as a high level of evidence without the consideration of effect size.
- Researchers should assess clinical meaningfulness by considering participants and subgroups rather than only aggregate scores; we recommend simple approaches that rely on descriptive statistics, beginning with comparisons of percentages of participants who declined, stayed the same, and improved between the treatment and control groups.
- Researchers should scrutinize baseline values, consider measurement error, and assess the fragility of their results.

Conclusion

What is clear from this analysis is that the effects of the Well Elderly II study are not as straightforward as they appear at first glance. More important, these findings show how misleading statistically significant *p* values can be for intervention researchers and, perhaps more so, for those investigators citing intervention research that claims effectiveness. After all, one cannot be expected to spend weeks sorting through original data to reassess the effectiveness of published interventions.

The approaches we used only indirectly address clinical meaningfulness, and none of them alone solve the puzzle of whether changes were meaningful or an intervention was effective. However, it is important to note that all of the approaches promote a particular attitude during data analysis: *p* values based on aggregate changes should be the starting point for additional, descriptive analyses that zero in on participants and subgroups. These additional analyses

should consider measurement error, baseline scores, extreme improvers, and regression to the mean (e.g., by considering the distribution of change scores in the control group).

Before the analysis, though, investigators should be clear about intended treatment effects and mechanisms of change. The chosen assessments should, as much as possible, directly measure the predefined treatment effects. Expectations about how these effects operate in the population should be realistic: For some studies, such as pharmacological studies, small effects that reach most participants may indicate intervention success; for others, reaching only a limited number of participants may indicate a need to redesign the intervention or target only extreme values at baseline. Should the investigators choose to cast a wide net with a large number of assessments, it should be noted that under the null hypothesis of no change, statistical significance will be achieved 5% of the time by chance (i.e., at $\alpha = .05$).

References

Beaton, D. E., Bombardier, C., Katz, J. N., Wright, J. G., Wells, G., Boers, M., . . . Shea, B.; OMERACT MCID Working Group (2001). Looking for important change/differences in studies of responsiveness. *Journal of Rheumatology*, *28*, 400–405.

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error. *British Medical Journal*, *312*, 1654. https://doi.org/10.1136/bmj.312.7047.1654

Busija, L., Osborne, R. H., Nilsdotter, A., Buchbinder, R., & Roos, E. M. (2008). Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. *Health and Quality of Life Outcomes*, *6*, 55. https://doi.org/10.1186/1477-7525-6-55

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399. https://doi.org/10.17763/haer.48.3.t490261645281841

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., . . . Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: Results of the Well Elderly 2 randomised controlled trial. *Journal of Epidemiology and Community Health*, *66*, 782–790. https://doi.org/10.1136/jech.2009.099754

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., Jr., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *Spine Journal*, *7*, 541–546. https://doi.org/10.1016/j.spinee.2007.01.008

Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, *56*, 395–407. https://doi.org/10.1016/S0895-4356(03)00044-1

de Vet, H. C. W., Beckerman, H., Terwee, C. B., Terluin, B., & Bouter, L. M. (2006). Definition of clinical differences. *Journal of Rheumatology*, *33*, 434–435.

de Vet, H. C. W., & Terwee, C. B. (2010). The minimal detectable change should not replace the minimal important difference. *Journal of Clinical Epidemiology*, *63*, 804–805. https://doi.org/10.1016/j.jclinepi.2009.12.015

Farrar, J. T., Portenoy, R. K., Berlin, J. A., Kinman, J. L., & Strom, B. L. (2000). Defining the clinically important difference in pain outcome measures. *Pain*, *88*, 287–294. https://doi.org/10.1016/S0304-3959(00)00339-0

Freeman, P. R. (1993). The role of *p*-values in analysing trial results. *Statistics in Medicine*, *12*, 1443–1452. https://doi.org/10.1002/sim.4780121510

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The *p* value fallacy. *Annals of Internal Medicine*, *130*, 995–1004. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R.; Clinical Significance Consensus Meeting Group. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings*, *77*, 371–383. https://doi.org/10.4065/77.4.371

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, *10*, 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Inzitari, D., Simoni, M., Pracucci, G., Poggesi, A., Basile, A. M., Chabriat, H., . . . Pantoni, L.; LADIS Study Group. (2007). Risk of rapid global functional decline in elderly patients with severe cerebral age-related white matter changes: The LADIS study. *Archives of Internal Medicine*, *167*, 81–88. https://doi.org/10.1001/archinte.167.1.81

Jackson, J., Carlson, M., Mandel, D., Zemke, R., & Clark, F. (1998). Occupation in lifestyle redesign: The Well Elderly Study Occupational Therapy Program. *American Journal of Occupational Therapy*, *52*, 326–336. https://doi.org/10.5014/ajot.52.5.326

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352. https://doi.org/10.1016/S0005-7894(84)80002-7

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*, 407–415. https://doi.org/10.1016/0197-2456(89)90005-6

Johansson, A., & Björklund, A. (2016). The impact of occupational therapy and lifestyle interventions on older persons' health, well-being, and occupational adaptation. *Scandinavian Journal of Occupational Therapy*, *23*, 207–219. https://doi.org/10.3109/11038128.2015.1093544

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*, 1524–1529. https://doi.org/10.1097/00004583-200312000-00022

Lin, M., Lucas, H. C., Jr., & Shmueli, G. (2013). Too big to fail: Large samples and the *p*-value problem. *Information Systems Research*, *24*, 906–917. https://doi.org/10.1287/isre.2013.0480

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*, 304–305. https://doi.org/10.1037/h0025105

Mountain, G., Windle, G., Hind, D., Walters, S., Keertharuth, A., Chatters, R., . . . Roberts, J. (2017). A preventative lifestyle intervention for older adults (lifestyle matters): A randomised controlled trial. *Age and Ageing*, *46*, 627–634. https://doi.org/10.1093/ageing/afx021

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582–592. https://doi.org/10.1097/01.MLR.0000062554.74615.4C

Ohl, A., & Schelly, D. (2017). Beyond *p*-values: A case for clinical relevance. *British Journal of Occupational Therapy*, *80*, 752–755. https://doi.org/10.1177/0308022617735048

Paterniti, S., Verdier-Taillefer, M. H., Dufouil, C., & Alpérovitch, A. (2002). Depressive symptoms and cognitive decline in elderly people: Longitudinal study. *British Journal of Psychiatry*, *181*, 406–410. https://doi.org/10.1192/bjp.181.5.406

Portney, L. G., & Watkins, M. P. (2015). *Foundations of clinical research: Applications to practice*. Philadelphia: F.A. Davis.

Radloff, L. S. (1977). The CES–D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. https://doi.org/10.1177/014662167700100306

Rejas, J., Pardo, A., & Ruiz, M. A. (2008). Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *Journal of Clinical Epidemiology*, *61*, 350–356. https://doi.org/10.1016/j.jclinepi.2007.05.011

Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., & Kinlaw, A. (2011). The use of one- versus two-tailed tests to evaluate prevention programs. *Evaluation and the Health Professions*, *34*, 135–150. https://doi.org/10.1177/0163278710388178

Samsa, G., Edelman, D., Rothman, M. L., Williams, G. R., Lipscomb, J., & Matchar, D. (1999). Determining clinically important differences in health status measures: A general approach with illustration to the Health Utilities Index Mark II. *PharmacoEconomics*, *15*, 141–155. https://doi.org/10.2165/00019053-199915020-00003

Terwee, C. B., Roorda, L. D., Dekker, J., Bierma-Zeinstra, S. M., Peat, G., Jordan, K. P., . . . de Vet, H. C. (2010). Mind the MIC: Large variation among populations and methods. *Journal of Clinical Epidemiology*, *63*, 524–534. https://doi.org/10.1016/j.jclinepi.2009.08.010

Turner, D., Schünemann, H. J., Griffith, L. E., Beaton, D. E., Griffiths, A. M., Critch, J. N., & Guyatt, G. H. (2010). The minimal detectable change cannot reliably replace the minimal important difference. *Journal of Clinical Epidemiology*, *63*, 28–36. https://doi.org/10.1016/j.jclinepi.2009.01.024

van Gelder, B. M., Tijhuis, M. A. R., Kalmijn, S., Giampaoli, S., Nissinen, A., & Kromhout, D. (2004). Physical activity in relation to cognitive decline in elderly men: The FINE Study. *Neurology*, *63*, 2316–2321. https://doi.org/10.1212/01.WNL.0000147474.29994.35

Walsh, M., Srinathan, S. K., McAuley, D. F., Mrkobrada, M., Levine, O., Ribic, C., . . . Devereaux, P. J. (2014). The statistical significance of randomized controlled trial results is frequently fragile: A case for a fragility index. *Journal of Clinical Epidemiology*, *67*, 622–628. https://doi.org/10.1016/j.jclinepi.2013.10.019

Ware, J. E., Jr. (2000). SF-36 Health Survey update. *Spine*, *25*, 3130–3139. https://doi.org/10.1097/00007632-200012150-00008

Ware, J. E., Kosinski, M., & Keller, S. K. (1994). *SF-36 Physical and Mental Health Summaries Scales: A user's manual*. Boston: Health Institute.

Weir, J. P. (2005). Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*, 231–240.

Wood, V., Wylie, M. L., & Sheafor, B. (1969). An analysis of a short self-report measure of life satisfaction: Correlation with rater judgments. *Journal of Gerontology*, *24*, 465–469. https://doi.org/10.1093/geronj/24.4.465

Wyrwich, K. W., Bullinger, M., Aaronson, N., Hays, R. D., Patrick, D. L., & Symonds, T.; Clinical Significance Consensus Meeting Group. (2005). Estimating clinically significant differences in quality of life outcomes. *Quality of Life Research*, *14*, 285–295. https://doi.org/10.1007/s11136-004-0705-2

Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., & Wolinsky, F. D. (1999). Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. *Medical Care*, *37*, 469–478. https://doi.org/10.1097/00005650-199905000-00006

David Schelly, PhD, is Assistant Professor, Occupational Therapy Department, Clarkson University, Potsdam, NY; dschelly@clarkson.edu

Alisha Ohl, PhD, is Assistant Professor, Occupational Therapy Department, Clarkson University, Potsdam, NY.

Acknowledgments

We thank the reviewers for their helpful feedback, as well as several individuals who were kind enough to comment on early drafts. Thank you to Aaron Eakman, Cathy Schelly, Jerry Vaske, and Margaret Kaplan.

Citation: Schelly, D., & Ohl, A. (2019). Examining clinical meaningfulness in randomized controlled trials: Revisiting the Well Elderly II. *American Journal of Occupational Therapy, 73,* 7301205120. https://doi.org/10.5014/ajot.2019.030874